ADHD-Lang: A Large-Scale Social Media Dataset for Verbal Behavior and Digital Phenotyping in Adult ADHD

Daniel Wiechmann[1,2], Elma Kerz[2], Edward Kempa[3] and Yu Qiao[2]
[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2] Exaia Technologies GmbH, Germany
[3]Department of Computer Science, University of Florida, USA
d.wiechmann@[uva.nl|exaia-tech.com], e.kerz@exaia-tech.com,
kempaedward@ufl.edu, y.qiao@exaia-tech.com

Abstract

We introduce ADHD-Lang, a large-scale language resource derived from Reddit to advance computational phenotyping of adult ADHD. The corpus is constructed using a high-precision self-disclosure pattern to confirm ADHD diagnoses and a matched control cohort, comprising 12,070 ADHD users (317,073 posts; 2.83M sentences) and 12,070 controls (174,765 posts; 1.27M sentences). In releasing ADHD-Lang to the research community, we also provide the first comprehensive baseline results, systematically examining the accuracy–transparency trade-off across three model families: (1) interpretable shallow machine learning models trained on clinically meaningful, expert-engineered language biomarkers; (2) a deep BiLSTM network trained on the same feature representations to capture temporal dynamics across users' posts; and (3) black-box transformer-based models (BERT, RoBERTa, MentalRoBERTa) leveraging contextual embeddings—non-interpretable, high-dimensional representations. ADHD-Lang is released as a standardized benchmark to promote reproducible research and accelerate progress toward digital verbal-behavior phenotyping for adult ADHD.

Keywords: language resources, ADHD, mental health, social media, interpretability, transformers, benchmarking

1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is one of the most prevalent mental health disorders, characterized by persistent patterns of inattention, impulsivity, and hyperactivity, as defined in the *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition, Text Revision (DSM-5–TR; American Psychiatric Association, 2022). Once considered a childhood condition, ADHD is now recognized as a lifespan disorder, with symptoms enduring into adulthood for a substantial proportion of individuals (Faraone et al., 2021). Epidemiological evidence indicates an adult ADHD prevalence of approximately 2–5%, with significant underdiagnosis and delayed recognition worldwide (Ayano et al., 2023). Despite its prevalence, ADHD in adults remains comparatively under-investigated, largely due to diagnostic biases favoring pediatric populations and the historical absence of validated adult diagnostic criteria (Posner et al., 2020). Adult ADHD imposes a substantial economic burden. In the United States alone, the annual societal cost is estimated at $122.8 billion, driven predominantly by unemployment, lost productivity, and excess healthcare expenditures (Schein et al., 2022). Beyond its economic impact, ADHD is associated with substantial disruption across multiple life domains. The disorder is linked to impairments in academic achievement, social relationships, and adaptive functioning, with lasting adverse effects on educational attainment, career progression, interpersonal connections, and overall mental health (Shaw et al., 2012; Faraone et al., 2021). It is also associated with elevated risks of chronic physical health conditions–most notably obesity–collectively constituting a substantial personal, familial, and socioeconomic burden (Cortese et al., 2016; French et al., 2024).

Current ADHD screening and diagnostic methods in adults primarily rely on the DSM-5 (American Psychiatric Association, 2022) or ICD-11 (World Health Organization, 2019), applied through structured clinical interviews and standardized behavioral rating scales regarded as the gold standard for assessment (Posner et al., 2020; Marshall et al., 2021). Among the most commonly used self-report rating scales are the Adult ADHD Self-Report Scale (ASRS-v1.1) (Kessler et al., 2005) and the Barkley Adult ADHD Rating Scale–IV (BAARS-IV) (Barkley, 2011), which quantify symptom severity and functional impairment across daily life domains. This current gold standard faces significant limitations. Its reliance on retrospective self-report and subjective evaluation renders it vulnerable to recall bias, response distortion, variability in subjective perception, and episodic memory limitations (Podsakoff et al., 2003; Brevik et al., 2020). For example, cognitive heuristics such as the peak–end rule can distort recall by causing disproportionate weighting

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7279–7291
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

of the most intense and recent symptoms during self-report (Horwitz et al., 2023). Adult ADHD assessment and diagnosis are further complicated by the fact that the DSM-5 and ICD-11 diagnostic frameworks were originally developed for pediatric populations, emphasizing overt hyperactivity and impulsivity. In contrast, the clinical presentation of adult ADHD is more complex and heterogeneous, encompassing internalized hyperactivity, disorganization, and emotional dysregulation (Asherson et al., 2016; Maltezos et al., 2020). Compared to children, adults typically possess higher verbal abilities and employ compensatory or camouflaging strategies that mask underlying attentional and self-regulatory deficits, further complicating diagnostic accuracy (van der Putten et al., 2024). These limitations underscore an urgent need for more nuanced and multifaceted assessment approaches that integrate behavioral, cognitive, and digital phenotyping strategies to derive ecologically valid, quantifiable, and objective indicators from complementary data sources (see Section 2: Related Work on emerging digital phenotyping modalities).

Against this background, social media mining has emerged as a promising avenue for mental health research, offering ecologically valid, large-scale, and computationally derived data that capture behavioral signals shown to be informative across diverse mental health prediction and assessment tasks (Chancellor and De Choudhury, 2020; Correia et al., 2020; Garg, 2023). The ubiquity of social media platforms has transformed how individuals externalize internal states: users routinely share their interests (Zarrinkalam et al., 2018), daily experiences (Saha et al., 2022), and emotions (Prieto et al., 2014), generating abundant, naturally occurring data that reflect everyday self-expression and affective dynamics (Fassi et al., 2025; Baydili et al., 2025). These digital footprints complement conventional assessment approaches by revealing subtle behavioral and linguistic markers often inaccessible in structured clinical contexts (Zarate et al., 2022). Building on this foundation, research has increasingly focused on *digital phenotyping*—the computational translation of digital behavior into measurable indicators of mental health (Jain et al., 2015; Insel, 2018). Advances in artificial intelligence (AI), particularly natural language processing (NLP) and machine or deep learning (ML/DL), now enable the modeling of mental health–relevant features from large-scale, unstructured social media data. Leveraging user-generated content through these approaches, researchers are developing unobtrusive, scalable techniques for identifying risk factors, supporting population-level screening, and informing personalized mental health care. Together, these computational frameworks are shaping a new paradigm in psychiatry—one in which digital phenotyping augments traditional diagnostics to enable earlier detection, continuous monitoring, and personalized, adaptive interventions (Fischer et al., 2025). A critical driver of scientific progress in the research area of computational psychiatry has been the emergence of benchmark datasets that enable standardized evaluation, replication, and cross-model comparison (Paulus et al., 2016). Yet, existing language resources primarily focus on depression, anxiety, or suicidality (Garg, 2023; Ríssola et al., 2021), leaving adult ADHD markedly underrepresented. This absence limits systematic exploration of how ADHD-specific linguistic and behavioral signatures manifest in naturalistic discourse. Building on these considerations, the present paper makes three main contributions:

1. **Resource:** We introduce ADHD-Lang, a large-scale Reddit corpus for adult ADHD research with ground truth labels for positively annotated (diagnosed) users. Both ADHD and control users were identified following established protocols and procedures reimplemented from prior social media mental health benchmark datasets to ensure methodological consistency. ADHD-Lang will be released for non-commercial research use to promote transparency, reproducibility, and scientific progress in computational psychiatry.

2. **Baselines: Accuracy–Transparency Trade-offs:** We provide comprehensive baseline results on ADHD-Lang using three model families: (1) interpretable shallow machine learning (ML) models trained on clinically meaningful, expert-engineered linguistic features; (2) a deep BiLSTM network trained on the same feature representations to capture temporal dynamics across users' posts; and (3) black-box transformer-based models (BERT, RoBERTa, MentalRoBERTa) leveraging high-dimensional representations.

3. **Interpretability: Linguistic Signatures of Adult ADHD:** We conduct a feature importance analysis on the best-performing shallow ML model to identify the linguistic features most indicative of ADHD. Extending beyond the lexicon-based feature paradigms that have dominated prior ADHD research, we integrate content-agnostic metrics of linguistic sophistication, structural complexity, and communicative precision, grounded in multidisciplinary cognitive and behavioral science.

2. Related Work

As outlined in the Introduction, traditional ADHD diagnostic practices grounded in clinical interviews and DSM-5 or ICD-11 rating scales face well-documented limitations in objectivity and ecological validity. Recent advances in neuroimaging and electrophysiological methods, such as magnetic resonance imaging (MRI) and electroencephalography (EEG), have contributed to the development of objective clinical biomarkers for ADHD (Cao et al., 2023). While these clinic-based modalities have significantly advanced the objectivity and biological grounding of ADHD diagnostics, they remain resource-intensive, expert-dependent, and limited in scalability (Sibley, 2021; Mulraney et al., 2022). To extend this progress beyond specialized clinical contexts, digital phenotyping has emerged as a complementary paradigm for identifying behavioral biomarkers from naturalistic, real-world data. These approaches include wearable eye-tracking devices, revealing atypical visual attention dynamics in adults with ADHD (Jiang et al., 2024); wearables equipped with accelerometers, capturing real-world sleep and physical activity patterns predictive of ADHD (Rahman, 2025; Olinic et al., 2025); and voice biomarkers, reflecting prosodic and articulatory patterns linked to executive function deficits (von Polier et al., 2025). Collectively, these modalities demonstrate the feasibility of digital biomarkers for ADHD outside clinical settings.

In addition to sensor- and task-based approaches, a growing body of work highlights social media mining as a complementary modality for digital behavioral phenotyping (see reviews in the Introduction). Platforms such as Twitter and Reddit offer large-scale, naturally occurring language datasets with unprecedented sample sizes and temporal granularity compared to traditional lab- or clinic-based studies. While fewer studies have examined ADHD than other mental health conditions, existing work provides converging evidence that social media–derived digital footprints can distinguish adults with ADHD from neurotypical controls. Guntuku et al. (2017) analyzed posts from 1,399 Twitter users with self-reported ADHD and matched controls, finding that ADHD users posted more frequently and used more negative emotion, anger, and anxiety words, alongside self-referential and tentative language. Topic modeling revealed themes of disorganization, mental fatigue, failure, and stimulant use, aligning with clinical symptoms of inattention and emotional dysregulation. Cohan et al. (2018) examined 10,098 Reddit users, showing that individuals with ADHD produced more informal and impulsive language, expressed greater emotional intensity, and more often referenced attention difficulties and cognitive frustration than matched controls. Chen et al. (2023b) analyzed Twitter data from 3,135 ADHD users and 3,223 controls, replicating earlier findings of negative affect, self-referential language, and cognitive-effort themes, while additionally identifying increased nocturnal posting and stronger emotional intensity, interpreted as indicators of affective volatility and circadian disruption. Together, these studies provide the first systematic evidence that language patterns in naturalistic online communication can serve as a basis for developing digital verbal behavior biomarkers of ADHD.

3. Dataset Construction

In this section, we introduce the construction of the ADHD-Lang dataset. Section 3.1 describes the data collection procedure, including the systematic identification of users who publicly self-reported an official ADHD diagnosis and the selection of matched control users. Section 3.2 outlines the preprocessing pipeline, which adheres to established standards in social-media–based mental health research by removing platform-specific artifacts, normalizing text representations, and filtering superficial linguistic cues that might trivially distinguish diagnostic groups. Section 3.3 presents corpus statistics and the dataset split prepared for benchmarking and reproducible evaluation.

3.1. Data Collection

We developed the ADHD-Lang corpus by reimplementing and refining the data collection procedure of the Self-reported Mental Health Diagnoses (SMHD) corpus (Cohan et al., 2018), a widely used Reddit benchmark that includes multiple mental health conditions but is limited to data collected before 2018. To obtain a more recent and condition-specific dataset, we adapted SMHD's high-precision identification framework to focus exclusively on attention-deficit/hyperactivity disorder (ADHD). Reddit users and their posts were obtained from the Arctic Shift Project[1], a compliant and sustainable archive providing public access to Reddit data. This choice was motivated by the 2023 Reddit API policy changes, which introduced usage fees and restricted third-party access, rendering prior sources such as Pushshift largely inaccessible. Arctic Shift offers a sustainable alternative through large-scale data dumps and a limited API aligned with current platform policies. Reddit[2], an anonymous discussion platform structured around topic-specific communities known as *subreddits*, provides longer, context-rich user-generated text compared with microblogging platforms such as Twitter (now X)[3], enabling the extraction of high-resolution digital markers of verbal behavior relevant to computational phenotyping.

[1] https://arctic-shift.org
[2] https://www.reddit.com
[3] https://x.com

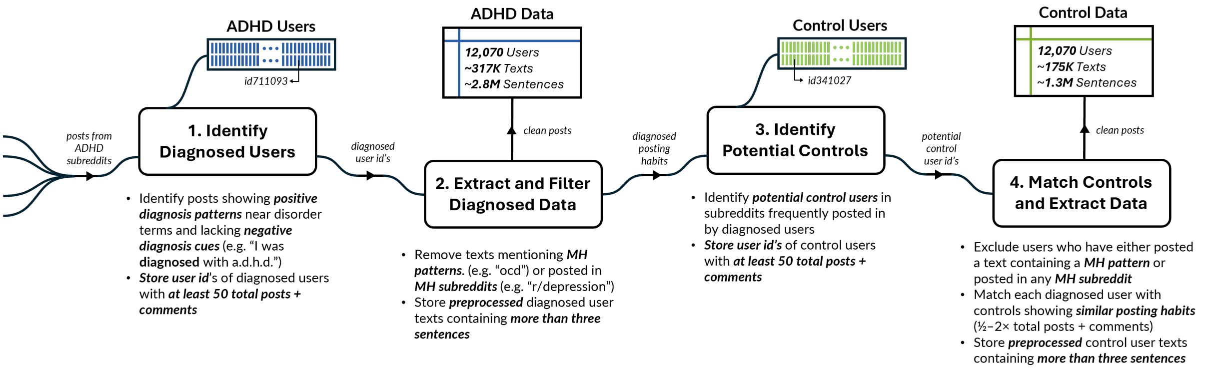

Figure 1: ADHD-Lang cohort construction and ecological sampling. We re-implemented the SMHD high-precision self-disclosure procedure on Reddit. A user is labeled ADHD only when a diagnosis phrase (e.g., "diagnosed with," "officially diagnosed") co-occurs with an ADHD keyword within $\leq$40 characters in the same post. Negations and hypotheticals are filtered (e.g., "never clinically diagnosed," "I think I might have...," "going to get a diagnosis"). After identifying diagnosed users, we collect their posts from non-mental-health subreddits and exclude mental-health communities, yielding more ecologically valid samples of everyday verbal behavior beyond r/ADHD.

Diagnosed ADHD users were identified using a high-precision pattern-matching procedure designed to capture only individuals who publicly disclosed an official diagnosis. Each detection pattern combined (1) diagnostic expressions (e.g., "I was diagnosed with") and (2) ADHD-related keywords, including common spelling and orthographic variants such as "adhd," "a.d.h.d.," or "add." A user was labeled as ADHD if a keyword occurred within 40 characters of a diagnostic expression, and all such disclosures were searched within mental-health-related subreddits where formal diagnoses are typically discussed. For each identified user, we then retrieved their complete posting history and applied a two-stage filtering process.

First, posts originating from mental-health-related or support-oriented subreddits were removed, covering 110 communities in total (the 78 used in SMHD plus 32 newer ones introduced after 2020). Second, posts containing explicit mental-health vocabulary (e.g., "diagnosis," "therapy," "depression") were filtered out to prevent label leakage and ensure that features were derived solely from general, non–mental-health discourse. Users with fewer than 50 remaining posts after filtering were excluded to maintain sufficient linguistic data for analysis. Control users were randomly sampled to match the ADHD cohort in posting volume and subreddit participation but were excluded if they mentioned ADHD or engaged in mental-health discussions. Manual validation of 500 detected instances confirmed a labeling precision of 97%, exceeding the 93% reported for SMHD. All detection patterns, keyword variants, subreddit lists, and filtering resources are publicly available in the project's GitHub repository.[4]

3.2. Data Preprocessing

The preprocessing pipeline standardized the text and removed platform-specific noise (e.g., URLs, markup remnants, and link/media placeholders) to improve consistency across samples. Implementation was carried out in Python (v3.10) using regular expressions to systematically clean and normalize text. All text was lowercased, and hyperlinks, HTML tags, HTML entities (including zero-width spaces), and user mentions were removed. Hashtags were normalized by deleting the hash symbol while retaining the lexical token (e.g., "#focus" $\rightarrow$ "focus"). Non-alphanumeric characters were stripped, except for a restricted set of punctuation marks (periods, commas, apostrophes, exclamation marks, question marks, semicolons, and hyphens). Consecutive whitespace was collapsed, and leading or trailing spaces were trimmed. Sentence boundaries were determined using punctuation cues (periods, exclamation marks, question marks), and sentence fragments shorter than two characters were discarded. Repeated or redundant punctuation (e.g., multiple exclamation marks or ellipses) was reduced to a single instance, and missing terminal punctuation was appended when necessary to ensure consistent sentence termination. Cleaned sentences were concatenated into normalized text strings while preserving sentence boundaries. The complete Python implementation is available in the project's GitHub repository to ensure transparency and reproducibility.

[4] https://github.com/Eddi005/LREC_2026_ADHD.

3.3. Data Analysis

The final ADHD-Lang dataset comprises 12,070 users who publicly disclosed an official ADHD diagnosis and an equal number of matched control users. Table 1 summarizes the dataset composition after all preprocessing and filtering steps. To support reproducible modeling and standardized benchmarking, an official user-level train/validation/test split is provided. Users were partitioned by stratified random sampling (80/10/10) to preserve class balance between ADHD and control groups. All partitions are strictly user-disjoint, ensuring that each individual and their complete posting history appear in only one subset. This design eliminates information leakage across data partitions and guarantees that model evaluation reflects generalization to unseen users.

4. Experimental Setup

This section describes the experimental framework developed to evaluate automated ADHD detection within the broader context of social media mining for mental health. The task is formulated as a binary user-level classification problem, in which models predict diagnostic group membership from preprocessed social media text. The experiments are designed to provide a comparative assessment of approaches along the interpretability spectrum, quantifying trade-offs between machine learning and deep learning methods. Section 4.1 outlines the evaluation metrics used for benchmarking, and Section 4.2 details the baseline architectures and configurations.

4.1. Evaluation Metrics

Model performance was assessed using standard metrics for binary text classification. The primary evaluation metric was the F1-score for the ADHD (positive) class, calculated as the harmonic mean of precision and recall to balance false positives and false negatives. For comprehensive benchmarking and comparison with prior work in social media–based mental health research, we also report precision, recall, and overall accuracy.

4.2. Baseline Models

We evaluated ADHD-Lang using three model families: (1) interpretable shallow machine-learning models—*Logistic Regression* (Hosmer et al., 2013), *Support Vector Machines* (Cortes and Vapnik, 1995), *Random Forests* (Breiman, 2001), and *Gradient Boosting* (Friedman, 2001)—trained on clinically meaningful, expert-engineered linguistic features; (2) a deep *Bidirectional Long Short-Term Memory* (BiLSTM) network (Schuster and Paliwal, 1997) trained on the same feature representations to capture temporal dependencies across users' posts; and (3) transformer-based models—*BERT* (Devlin et al., 2019), *RoBERTa* (Liu et al., 2019), and *MentalRoBERTa* (Ji et al., 2022)—leveraging contextual embeddings.

Together, these baselines illustrate the trade-offs between interpretability and accuracy in high-stakes domains such as healthcare, where transparency is as critical as predictive performance (Kerz et al., 2023; Ding et al., 2025). All models were implemented in Python 3, using *pandas* for data handling, *scikit-learn* (Pedregosa et al., 2011) for machine-learning algorithms, *PyTorch* (Paszke et al., 2019) for deep-learning architectures, and the *Transformers* library (Wolf et al., 2020) for pretrained encoders. All hyperparameters were tuned on the validation set, with the final configurations summarized in Table 2. All code and documentation are available in the project's GitLab repository, ensuring full reproducibility of the ADHD-Lang baseline framework.

Model families (1) and (2) were trained on quantitative metrics extracted with **EXAIA CYMO v1.0.0-beta**[5], a proprietary end-to-end text-mining and analytics platform for high-throughput data processing. CYMO employs a distributed architecture with parallel processing and dynamic resource allocation, enabling efficient annotation of large-scale datasets (491,838 text files in this study). Using a sliding-window approach, it computes sentence-level metrics that capture fine-grained distributions of indicators within text. Built on *spaCy* for core NLP operations—tokenization, lemmatization, POS tagging, dependency parsing, and sentence segmentation—CYMO provides a scalable, reproducible environment for transparent, high-resolution data analysis. The software is proprietary but freely accessible for research use, with public documentation and tutorials available online. CYMO currently supports 344 quantitative metrics grounded in multidisciplinary research on how humans learn and process language across the lifespan. The framework integrates theoretical and empirical insights from cognitive, affective, and behavioural neuroscience, computational and psycholinguistics, and educational research. These metrics are organized into two main classes: The first class, referred to as *core metrics*, is content-agnostic and quantifies the structural organization and informational density of language, reflecting variation in cognitive effort and planning. It comprises seven categories: *Syntactic Complexity* (17 metrics), *Lexical Diversity* (15), *Lexical Density* (1), *Lexical Sophistication* (116), *Lexical Overlap* (54), and *Information-Theoretic Measures* (1). The second class, *lexicon-based metrics*,

[5] https://github.com/exaiatech/cymo-tutorial

User Group	Users	Texts	Sentences	Total Words	Mean Texts per User	Mean Words per Text
Diagnosed	12,070	317,073	2,830,661	45,343,408	26.27	143.01
Control	12,070	174,765	1,266,155	18,021,569	14.48	103.12

Table 1: Statistics of the ADHD-Lang dataset.

Model	Hyperparameter Settings
LR	Regularization strength $C = 1$; L2 penalty
SVM	$C = 1$; linear kernel; class weight = balanced
RF	500 trees; max depth = 12; min samples split = 2
GB	500 estimators; learning rate = 0.05; max leaves = 128
BiLSTM	Hidden units per direction = 256; 3 layers; dropout = 0.2; learning rate = 0.001
TRF	Learning rate = 2×10^{-5}; weight decay = 10^{-5}; batch size = 16; epochs = 5

Table 2: Hyperparameter settings used in the ADHD-Lang experiments. Abbreviations: LR = Logistic Regression; SVM = Support Vector Machine; RF = Random Forest; GB = Gradient Boosting; BiLSTM = Bidirectional Long Short-Term Memory; TRF = Transformer-based models.

follows a closed-vocabulary approach and captures semantic, affective, and functional dimensions of word use. These include *LexEmo* (28 emotion metrics derived from the revised Hourglass Model; Susanto et al. 2020), *LexGram* (83 grammatical-pattern metrics describing function-word usage), and *LexTopic* (15 topical domains such as Health, Relationships, and Science). A full description of all 344 metrics, including definitions and computational formulas, is available in the project's GitLab repository.

In the context of the present study, which investigates the verbal behavior of adults with ADHD, the core metrics were used to explore hypotheses derived from Barkley's (1997) unified model of ADHD. This model attributes the disorder's cognitive and behavioral symptoms to deficits in executive function, inhibitory control, and self-regulation. Such impairments are hypothesized to manifest in text as less precise, more generic, and more repetitive language use—characterized by greater lexical overlap, lower lexical sophistication, and reduced structural complexity. For the lexicon-based (closed-vocabulary) metrics, our analyses test whether previously reported linguistic patterns in adults with ADHD—such as increased self-referential expression and elevated use of negative emotional language linked to emotional dysregulation—replicate and extend in our dataset. Beyond replication, we broaden the affective scope by moving past Ekman's six basic emotions toward a more differentiated taxonomy of emotional states, as operationalized in the emotion metrics described above. This expanded framework, together with the extended topical domains, enables a more nuanced and comprehensive characterization of affective and semantic variation in ADHD-related language use.

Model families (1) and (2) operationalized these hypotheses at different representational levels. The first family used text-level aggregates of all CYMO-derived metrics to model users' overall linguistic profiles, emphasizing global patterns of lexical and structural organization. The second, a bidirectional long short-term memory (BiLSTM) network, was trained on sentence-level metric sequences to capture temporal dependencies and within-user variability in linguistic behavior across posts.

5. Results and Discussion

Table 3 summarizes the performance of the eight binary classification models across standard evaluation metrics, including accuracy, precision, recall, and F_1-score.

Model	Acc.	P	R	F1
LR	74.21	73.91	74.86	74.38
RF	76.12	74.49	79.45	**76.89**
SVM	76.47	75.91	77.55	76.72
GB	76.52	75.33	78.46	76.86
BiLSTM	76.33	74.83	79.33	**77.02**
BERT	74.46	73.67	76.14	74.88
RoBERTa	75.72	75.14	76.88	76.00
MRoBERTa	76.01	74.51	79.08	**76.73**

Table 3: Performance metrics for binary ADHD classification. Accuracy (Acc.), Precision (P), Recall (R), and F1-score (F1) are reported for the ADHD class (%).

Model Family 1: Shallow machine-learning models. Among the interpretable classifiers, the Random Forest (RF) achieved the highest F_1-score (76.89%), followed closely by Gradient Boosting (GB, 76.86%) and SVM (76.72%). Logistic Regression (LR) performed lowest (74.38%). The similar F_1-scores of RF, GB, and the (linear-kernel) SVM suggest that most of the predictive signal is already well captured by the chosen feature representations, while the tree-based models may additionally benefit from feature interactions, yielding a consistent improvement of about 2.5–2.7% over the linear baseline. **Model Family 2: Deep BiLSTM model.** The BiLSTM achieved the highest F_1-score overall (77.02%), improving upon the best shallow model (RF) by 0.13 percentage points. Although the difference is small, it reflects a measurable benefit from processing sentence-level feature sequences, which preserve variation across discourse segments instead of relying solely on text-level aggregates. **Model Family 3: Transformer-based models.** Among the transformer architectures, MentalRoBERTa (M-RoBERTa) reached the best F_1-score (76.73%), outperforming RoBERTa (76.00%) and BERT (74.88%). The results indicate that domain adaptation enhances transformer performance, as M-RoBERTa—pretrained on mental health–related text—generalized better to ADHD language than general-purpose models. Nevertheless, the gains remained modest, suggesting that task-specific fine-tuning alone cannot fully capture the linguistic markers of ADHD without domain-informed feature representations. **Comparative assessment across model families.** Performance across model families was generally comparable, with F_1-scores ranging between 76–77%. The Random Forest (RF) emerged as the most balanced and interpretable model, achieving competitive performance relative to both the BiLSTM and transformer-based architectures. Given its strong accuracy and transparent decision structure, subsequent analyses focus on the RF model to examine which linguistic features most strongly contribute to ADHD classification and to provide interpretable insights into the underlying language markers. These findings suggest that interpretability and predictive accuracy are not mutually exclusive. The interpretable models in Family 1 (shallow machine learning) outperformed the transformer-based models in Family 3, demonstrating that transparent, linguistically grounded systems can achieve high predictive performance without reliance on opaque "black-box" architectures. More broadly, this work contributes to an emerging body of evidence indicating that the trade-off between interpretability and accuracy is less pronounced than often assumed. Similar results have been reported for other mental health conditions—including depression, bipolar disorder, anxiety, and personality disorders—highlighting the importance of interpretable modeling approaches for understanding and diagnosing psychiatric disorders (Kerz et al., 2023; Ding et al., 2025).

Building on these results, we next examine feature importance within the Random Forest model to identify the features most predictive of ADHD. This analysis provides an interpretable view of the verbal behavior patterns that drive model decisions and highlights which feature categories contribute most to classification performance. In doing so, it offers insights into the linguistic signatures underlying ADHD within a digital phenotyping framework.

Feature importance was quantified using the Mean Decrease in Impurity (MDI), a model-intrinsic measure that estimates each feature's contribution to reducing classification uncertainty. At each split in the decision trees, the reduction in Gini impurity attributed to the selected variable was computed and weighted by the proportion of samples reaching that node. These weighted impurity reductions were then averaged across all trees in the ensemble to yield a global estimate of each feature's relative importance.

Figure 2 presents the 20 most discriminative metrics ranked by Mean Decrease in Impurity (MDI). The highest-ranking features spanned both major classes of linguistic measures introduced in Section 4.2. Among these, emotion-related metrics were most frequent (9), followed by lexical overlap (6), grammatical (2), lexical sophistication (2), and topical (1). Collectively, these results show that the most predictive dimensions integrate core, content-agnostic markers of structural and lexical organization with lexicon-based features capturing affective, grammatical, and thematic aspects of language use.

As outlined in Section 4.2, our feature selection and engineering were guided by hypotheses targeting core, content-agnostic metrics, a dimension largely neglected in prior linguistic research on adult ADHD. These metrics were expected to capture structural manifestations of executive dysfunction—specifically greater lexical overlap, lower lexical sophistication, and reduced structural complexity. Consistent with these expectations, individuals with ADHD showed markedly higher lexical overlap across sentences than controls. The six top-ranked overlap measures quantified the recurrence of identical lexical content within or between utterances, indexing repetitiveness in discourse. Two further high-ranking features—normalized bigram (2GNLF) and trigram frequencies (3GNLF)—indexed reliance on frequent, formulaic multiword sequences. Elevated values in ADHD therefore indicate a preference for generic, high-frequency phrasing and reduced contextual adaptability. Within theoretical accounts linking ADHD to deficits in executive control and self-regulation, this pattern reflects limitations in the maintenance and updating of lexical representations. Consequently, individuals with ADHD rely more heavily on familiar, low-information expressions, yielding reduced linguistic variability and informational density.

Turning to the lexicon-based (closed-vocabulary) metrics, the results largely replicate and extend previous findings on emotion-related language use in adult ADHD. Consistent with earlier studies by

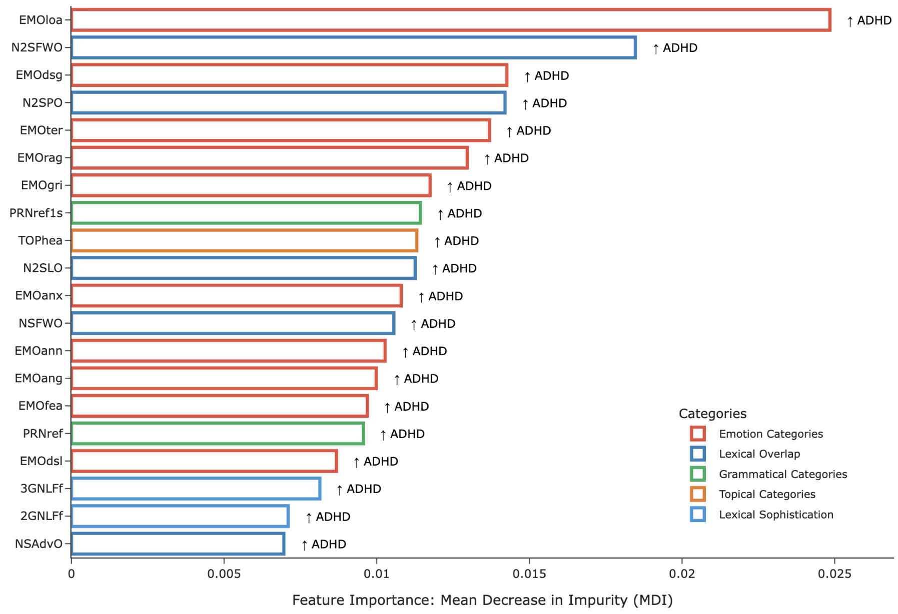

Figure 2: Top 20 linguistic features ranked by mean decrease in impurity (MDI) for adult ADHD classification using the Random Forest model. Emotion-related features (Loathing, Disgust, Anxiety, Anger, Fear) reflect the relative frequency of affective word use. Lexical Overlap features capture repetition across sentences, Grammatical features indicate pronoun usage, Topical features represent health-related vocabulary, and Lexical Sophistication features measure the normalized frequency of common n-grams. Arrows denote higher feature values in the ADHD group.

Guntuku et al. (2019) and Cohan et al. (2018), individuals with ADHD produced proportionally more negative emotion words—including terms related to anger, anxiety, fear, disgust, and loathing—than controls. This linguistic signature converges with the growing consensus that emotion dysregulation constitutes a core clinical dimension of ADHD. Although not yet a formal diagnostic criterion in the *DSM-5*, emotion dysregulation is increasingly recognized as central to the disorder's phenotype and is being actively considered for inclusion in future nosological revisions (Shaw et al., 2014; Thorell et al., 2020).

In addition, topical features (TOPhea) reflecting health-related vocabulary were more prevalent in ADHD language. Notably, this pattern persisted even when verbal behavior was drawn from non–mental-health-related subreddits (e.g., *soccer*), suggesting a persistent preoccupation with health-related themes across contexts (Chen et al., 2023a). Finally, language in adult ADHD was characterized by elevated use of reflexive and first-person singular pronouns (e.g., *myself*, *I*, *me*), echoing findings by Guntuku et al. (2019) and Cohan et al. (2018). Increased self-referential language has been interpreted as reflecting heightened self-focus, internalized emotional processing, or reduced social-attentional engagement—patterns consistent with broader profiles of executive and socio-emotional dysregulation in ADHD.

Together, these findings lay the groundwork for establishing digital phenotypes of ADHD derived from naturally occurring verbal behavior on social media. The present results provide a conceptual and empirical foundation for future work using the ADHD-Lang framework to identify, track, and potentially predict clinically relevant cognitive–affective states from language in ecologically valid, large-scale contexts.

6. Conclusion

We presented **ADHD-Lang**, a large-scale language resource for social media mining in mental health, constructed from Reddit to advance computational modeling and digital phenotyping of adult ADHD. The dataset was developed through a high-precision self-disclosure identification framework that ensures reliable ground-truth labels for diagnosed individuals and a rigorously matched control cohort. Combined with a transparent, standardized preprocessing pipeline, ADHD-Lang provides a high-quality and reproducible foundation for research on language-based markers of ADHD. To enable systematic and transparent benchmarking, we provide comprehensive baseline experiments across multiple model families, establishing reference points for future work examining the trade-offs between interpretability and predictive performance. These baselines are designed to support reproducible evaluation and to facilitate the development of explainable computational models in mental health NLP. By releasing ADHD-Lang with comprehensive documentation, ethical safeguards, and standardized evaluation splits, we aim to catalyze community-driven progress toward interpretable and clinically meaningful NLP models for mental health research. The resource is intended exclusively for academic research use under a controlled data-sharing agreement. Beyond ADHD, this work highlights the broader potential of interpretable computational linguistics to bridge digital behavior and clinical insight—enabling the development of transparent, ethically grounded tools for early detection, monitoring, and intervention across mental health conditions.

7. Ethical Considerations and Data Availability

This study relies on publicly available Reddit posts in full compliance with the platform's Privacy Policy[6] and Terms of Service. Reddit's policies explicitly allow third-party research access to public content through its API and similar technologies, provided users' privacy and data rights are respected. All user information was de-identified by replacing usernames with randomly assigned numeric identifiers. No attempts were made to contact users, link pseudonymous accounts across platforms, or recover personally identifiable information. Following established best practices for ethical and responsible social media and mental health NLP research (Hovy and Spruit, 2016; Losada et al., 2017; Cohan et al., 2018), the data collection and preprocessing process was designed to minimize any risk of re-identification or privacy breach.

This work is intended solely for scientific research and must not be used for individual diagnosis, risk assessment, or clinical decision-making. Automated models trained on social media data should not be deployed in clinical or public settings without rigorous validation, ethical oversight, and expert interpretation. The proposed methods are designed to complement—rather than replace—clinical expertise by providing interpretable, population-level insights into linguistic and cognitive markers of adult ADHD.

Data Availability. The processed dataset used in this study is available to qualified academic researchers upon request under a data use agreement, following the model of the SMHD benchmark introduced by Cohan et al. (2018). In line with ethical research standards, no raw Reddit data or user-generated text will be publicly released. Access is restricted to non-commercial research purposes, and requesters must provide proof of institutional affiliation and confirm approval (or an exemption) from their relevant ethics board/IRB for research involving sensitive mental-health–related content. The data use agreement further prohibits any attempt to re-identify individuals, link records to external profiles, or contact Reddit users, and it requires that derived models, analyses, and publications report only aggregated results and refrain from reproducing verbatim excerpts that could facilitate identification. We distribute only de-identified, preprocessed feature representations and dataset splits necessary to reproduce the analyses reported in this paper, and we provide code for the full preprocessing to support transparency and replicability without exposing user-generated text.

8. Limitations and Future Work

Despite its methodological rigor and scope, several limitations of the present work suggest directions for future research.

- **First,** ADHD-Lang may reflect selection biases inherent to individuals who publicly disclose an official ADHD diagnosis on Reddit. Such users may differ systematically from undiagnosed or non-disclosing individuals in demographic, psychological, or communicative characteristics, potentially shaping the linguistic patterns observed. Future work should therefore explore sampling strategies or complementary data sources to mitigate disclosure bias and evaluate the representativeness of ADHD-related language across broader populations.
- **Second,** the present study provides initial baseline models for benchmarking rather than optimized systems. While these results establish a reproducible foundation for model comparison, predictive accuracy can be further improved. Future research should investigate more advanced model architectures, including early- and late-fusion designs, stacking approaches, and hybrid ensemble architectures that capture complementary information across representational levels.
- **Third,** the present analysis of feature importance was limited to Gini impurity within a Random Forest framework, offering model-specific but global insights. Future work should expand interpretability using a broader range of explainable AI (XAI) techniques—spanning model-specific and model-agnostic, local and global, and ante-hoc and post-hoc approaches—to obtain a more comprehensive

[6] https://www.reddit.com/policies/privacy-policy

understanding of model mechanisms and feature relevance.

- **Fourth,** ADHD-Lang captures language use in a cross-sectional manner. Incorporating longitudinal data would allow for temporal modeling of intra-individual variation, enabling the study of how linguistic markers evolve over time.

Collectively, these extensions will advance ADHD-Lang toward a more accurate, interpretable, and temporally informed framework for digital behavioral phenotyping in adult ADHD.

9. References

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders (5th ed., text rev.; DSM-5–TR)*. American Psychiatric Publishing, Washington, DC.

Philip Asherson, Jan Buitelaar, Stephen V. Faraone, and Luís A. Rohde. 2016. Adult attention-deficit hyperactivity disorder: Key conceptual issues. *The Lancet Psychiatry*, 3(6):568–578.

Getinet Ayano, Kebede Yohannes, et al. 2023. Prevalence of attention deficit hyperactivity disorder in adults: Umbrella review of global evidence. *Psychiatry Research*, 328:115449.

Russell A. Barkley. 2011. *Barkley Adult ADHD Rating Scale–IV (BAARS–IV)*. Guilford Press, New York, NY.

A. Baydili, F. Tasci, and G. Tasci. 2025. Social media analytics for mental health: A comprehensive review of methods and applications. *Frontiers in Digital Health*, 7:122.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Erlend J. Brevik, Astri J. Lundervold, Jan Haavik, and Maj-Britt Posserud. 2020. Validity and accuracy of the adult ADHD self-report scale (ASRS) and the Wender Utah Rating Scale (WURS) symptom checklists in discriminating between adults with and without ADHD. *Brain and Behavior*, 10(6):e01605.

M. Cao, E. Martin, and X. Li. 2023. Machine learning in attention-deficit/hyperactivity disorder: new approaches toward understanding the neural mechanisms. *Translational Psychiatry*, 13(1):236.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive mental health research on social media: A review. *NPJ Digital Medicine*, 3(1):43.

Liuliu Chen, Jiwon Jeong, Bridgette Simpkins, and Emilio Ferrara. 2023a. Exploring the behavior of users with attention-deficit/hyperactivity disorder on Twitter: Comparative analysis of tweet content and user interactions. *Journal of Medical Internet Research*, 25:e43439.

Wei Chen, Han Wang, Zhiyuan Li, Jiayu Sun, and Yong Zhang. 2023b. Linguistic markers of ADHD on social media: Behavioral and temporal correlates from Twitter data. *Frontiers in Digital Health*, 5:117–130.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 167–178, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

R. B. Correia, I. B. Wood, J. Bollen, and L. M. Rocha. 2020. Mining social media data for biomedical signals and health insights. *NPJ Digital Medicine*, 3:47.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Samuele Cortese, Carlos R. Moreira-Maia, Delphine St Fleur, Concepción Morcillo-Peñalver, Luís A. Rohde, and Stephen V. Faraone. 2016. Association between ADHD and obesity: A systematic review and meta-analysis. *American Journal of Psychiatry*, 173(1):34–43.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Z. Ding, Z. Wang, Y. Zhang, Y. Cao, Y. Liu, X. Shen, and J. Dai. 2025. Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, 15(1):14497.

Stephen V. Faraone, Tobias Banaschewski, David Coghill, Yi Zheng, Joseph Biederman, Mark A. Bellgrove, Jeffrey H. Newcorn, Martin Gignac, Nouf M. Al Saud, Iris Manor, Luis A. Rohde, Li Yang, Samuele Cortese, et al. 2021. The World Federation of ADHD International Consensus Statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, 128:789–818.

L. Fassi, C. J. Ferguson, and A. Orben. 2025. Social media and mental health: Rethinking digital expression and psychological measurement. *Trends in Cognitive Sciences*. In press.

L. Fischer, P. A. Mann, M. H. Nguyen, S. Becker, S. Khodadadi, A. Schulz, S. E. Thanarajah, J. Repple, T. Hahn, A. Reif, A. Salamikhanshan, S. Kittel-Schneider, W. Rief, C. Mulert, S. G. Hofmann, U. Dannlowski, T. Kircher, F. P. Bernhard, and H. Jamalabadi. 2025. AI for mental health: Clinician expectations and priorities in computational psychiatry. *BMC Psychiatry*, 25(1):584.

Benjamin French, Gökhan Nalbant, Hannah Wright, Kapil Sayal, David Daley, Matthew J. Groom, Sarah Cassidy, and Clare L. Hall. 2024. The impacts associated with having ADHD: An umbrella review. *Frontiers in Psychiatry*, 15:1343314.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

M. Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

Sharath Chandra Guntuku, J. Russell Ramsay, Raina M. Merchant, Lyle H. Ungar, and colleagues. 2019. Language of ADHD in adults on social media. *Journal of Attention Disorders*, 23(12):1475–1485.

Sharath Chandra Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: An integrative review. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 43–54, Perth, Australia. International World Wide Web Conferences Steering Committee.

Ari M. Horwitz, T. Fehr, et al. 2023. Cognitive biases in self-reported mental health symptoms: Revisiting the peak–end rule. *Frontiers in Psychology*, 14:1122334.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*, 3rd edition. Wiley.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Thomas R. Insel. 2018. Digital phenotyping: a global tool for psychiatry. *World Psychiatry*, 17(3):276–277.

Sachin H. Jain, Brian W. Powers, Jared B. Hawkins, and John S. Brownstein. 2015. The digital phenotype. *Nature Biotechnology*, 33(5):462–463.

Shaoxiong Ji, Yi Zhang, Shuo Zhang, and Jürgen Pfeffer. 2022. MentalRoBERTa: A pretrained transformer for mental health text analysis. *arXiv preprint arXiv:2205.12700*.

Z. Jiang, A. Y. Chan, D. Lum, K. H. Wong, J. C. Leung, P. Ip, and I. C. Wong. 2024. Wearable signals for diagnosing attention-deficit/hyperactivity disorder in adolescents: A feasibility study. *JAACAP Open*. Advance online publication.

E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann. 2023. Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry*, 14:1219479.

Ronald C. Kessler, Lenard Adler, Matthew Ames, Olga Demler, Stephen V. Faraone, Eva Hiripi, et al. 2005. The world health organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, 35(2):245–256.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E. Losada, Fabio Crestani, and Javier Parapar. 2017. CLEF 2017 eRisk overview: Early risk prediction on the internet: Experimental foundations. In *Working Notes of CLEF 2017*.

Elemi Maltezos, Simon Whitwell, and Philip Asherson. 2020. Adhd in adults. In Niruj Agrawal and Mayur Bodani, editors, *Oxford Textbook of Neuropsychiatry*, pages 411–424. Oxford University Press, Oxford.

Paul Marshall, James Hoelzle, and Molly Nikolas. 2021. Diagnosing attention-deficit/hyperactivity disorder (ADHD) in young adults: A qualitative review of the utility of assessment measures and recommendations for improving the diagnostic process. *The Clinical Neuropsychologist*, 35(1):165–198.

M. Mulraney, G. Arrondo, H. Musullulu, I. Iturmendi-Sabater, S. Cortese, S. J. Westwood, et al. 2022. Systematic review and meta-analysis: Screening tools for attention-deficit/hyperactivity disorder in children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 61:982–996.

M. S. Olinic, R. Stretea, and C. Cherecheș. 2025. Wearables in ADHD: Monitoring and intervention—where are we now? *Diagnostics*, 15(18):2359.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, page 8024–8035.

Martin P. Paulus, Quentin J. Huys, and Tiago V. Maia. 2016. A roadmap for the development of computational models of psychiatric and neurological disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):386–392.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879–903.

Jonathan Posner, Guilherme V. Polanczyk, and Edmund Sonuga-Barke. 2020. Attention-deficit hyperactivity disorder. *The Lancet*, 395(10222):450–462.

V. M. Prieto, S. Matos, M. Alvarez, F. Cacheda, and J. L. Oliveira. 2014. Twitter: A good place to detect health conditions. *PLoS ONE*, 9(1):e86191.

M. M. Rahman. 2025. Unlocking the potential of wearable technology: Fitbit-derived measures for predicting ADHD in adolescents. *Frontiers in Child and Adolescent Psychiatry*, 4:1504323.

E. A. Ríssola, D. E. Losada, and F. Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(1):1–31.

Koustuv Saha, Ahmed Yousuf, Ryan L. Boyd, James W. Pennebaker, and Munmun De Choudhury. 2022. Social media discussions predict mental health consultations on college campuses. *Scientific Reports*, 12(1):1–13.

Jeffrey Schein, Douglas Faries, et al. 2022. Economic burden of attention-deficit/hyperactivity disorder among adults in the united states: A societal perspective. *Journal of Managed Care & Specialty Pharmacy*, 28(2):168–179.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Mark Shaw, Paul Hodgkins, Hervé Caci, Susan Young, Jens Kahle, Andrew G. Woods, and L. Eugene Arnold. 2012. A systematic review and analysis of long-term outcomes in attention-deficit hyperactivity disorder: Effects of treatment and non-treatment. *BMC Medicine*, 10:99.

Philip Shaw, Argyris Stringaris, Joel Nigg, and Ellen Leibenluft. 2014. Emotion dysregulation in attention-deficit/hyperactivity disorder. *American Journal of Psychiatry*, 171(3):276–293.

M. H. Sibley. 2021. Empirically-informed guidelines for first-time adult ADHD diagnosis. *Journal of Clinical and Experimental Neuropsychology*, 43:340–351.

Yohan Susanto, Erik Cambria, Dipankar Das, and Björn Schuller. 2020. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Lisa B. Thorell, Hanna Tilling, and Douglas Sjöwall. 2020. Emotion dysregulation in adult ADHD: Introducing the comprehensive emotion regulation inventory (CERI). *Journal of Clinical and Experimental Neuropsychology*, 42(7):747–758.

W. J. van der Putten, A. J. J. Mol, A. P. Groenman, T. A. Radhoe, C. Torenvliet, J. A. Agelink van Rentergem, and H. M. Geurts. 2024. Is camouflaging unique for autism? a comparison of camouflaging between adults with autism and ADHD. *Autism Research*, 17(4):812–823.

G. G. von Polier, E. Ahlers, J. Volkening, J. Langner, K. R. Patil, S. B. Eickhoff, and D. Langner. 2025. Exploring voice as a digital phenotype in adults with ADHD. *Scientific Reports*, 15(1):18076.

Thomas Wolf, Lysandre Début, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45.

World Health Organization. 2019. ICD-11 for mortality and morbidity statistics (version: 2019 april). https://icd.who.int/browse11/l-m/en. Accessed 2025-10-17.

D. Zarate, C. Hohmann, and C. Walsh. 2022. Digital traces of emotion: How language reflects psychological states on social media. *Journal of Affective Computing*, 13(4):547–561.

F. Zarrinkalam, M. Kahani, and A. Shakery. 2018. Mining user interests over microblogs: A review of the state of the art. *Social Network Analysis and Mining*, 8(1):1–23.