

A cross-dataset study on automatic detection of autism spectrum disorder from text data

Aleksander Wawer¹  | Izabela Chojnicka² | Justyna Sarzyńska-Wawer³ | Małgorzata Krawczyk³

¹Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland

²Department of Health and Rehabilitation Psychology, Faculty of Psychology, University of Warsaw, Warsaw, Poland

³Polish Academy of Sciences, Institute of Psychology, Warsaw, Poland

Correspondence

Aleksander Wawer, Polish Academy of Sciences, Institute of Computer Science, Jana Kazimierza 5, 01-248 Warsaw, Poland.
 Email: axw@ipipan.waw.pl

Funding information

National Science Centre (Poland), Grant/Award Numbers: 2020/39/D/HS6/00809, 2018/31/B/HS6/02848; Ministry of Science and Higher Education (Poland), Grant/Award Numbers: 501-D125-01-1250000 zlec.5011000231, 5011000228

Abstract

Objective: The goals of this article are as follows. First, to investigate the possibility of detecting autism spectrum disorder (ASD) from text data using the latest generation of machine learning tools. Second, to compare model performance on two datasets of transcribed statements, collected using two different diagnostic tools. Third, to investigate the feasibility of knowledge transfer between models trained on both datasets and check if data augmentation can help alleviate the problem of a small number of observations.

Method: We explore two techniques to detect ASD. The first one is based on fine-tuning HerBERT, a BERT-based, monolingual deep transformer neural network. The second one uses the newest, multipurpose text embeddings from OpenAI and a classifier. We apply the methods to two separate datasets of transcribed statements, collected using two different diagnostic tools: thought, language, and communication (TLC) and autism diagnosis observation schedule-2 (ADOS-2). We conducted several cross-dataset experiments in both a zero-shot setting and a setting where models are pretrained on one dataset and then training continues on another to test the possibility of knowledge transfer.

Results: Unlike previous studies, the models we tested obtained average results on ADOS-2 data but reached very good performance of the models in TLC. We did not observe any benefits from knowledge transfer between datasets. We observed relatively poor performance of models trained on augmented data and hypothesize that data augmentation by back translation obfuscates autism-specific signals.

Conclusion: The quality of machine learning models that detect ASD from text data is improving, but model results are dependent on the type of input data or diagnostic tool.

1 | INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition with a lifelong impact on social relations, independence, and everyday functioning. We will use the term autism along with the acronym ASD to refer to the autism spectrum condition. The prevalence of ASD, estimated at 2.8% in the United States and 1% worldwide, has consistently risen over the past 20 years.^{1–3} Autism has become a more recognizable social challenge. The diagnostic process of autism is complex and demanding, should take into account multiple sources of information, and involve various specialists. Assessment relies on behavioral factors that vary greatly from person to person. In many places around the world access to specialists is limited, and waiting times for specialist consultations and therapy are long. Machine learning has the potential to provide tools to support clinicians in decision-making and serve as a great equalizer in terms of access to diagnosis. Many studies have explored various approaches toward improved digital screening, diagnostics, and digital therapies for children with autism.^{4,5} Numerous data modalities encode behavior-rich information for machine-learning autism phenotyping. Among them are data derived from questionnaires—once in paper and pencil format, but today increasingly in digital form (e.g., Reference 6). Another source of data is video recordings of interactions, typically involving infants and toddlers (e.g., Reference 7). Studies using eye-tracking provide a specific type of data, namely gaze trajectories and visual attention, which have also been utilized for autism detection using computational models (e.g., Reference 8).

Another source of data for machine learning is textual data based on spoken or written language. Autistic individuals constitute a highly heterogeneous group in terms of language and communication abilities across all language subsystems, including pragmatics, semantics, grammar, syntax, morphology, and phonology, as well as verbal and non-verbal communication.⁹ Some individuals on the autism spectrum are non-speaking, some experience delays and disorders in language development, while at the other end of the spectrum are individuals who fluently use speech and display structural language skills within the typical range. At the same time, characteristics of speech and communication are key symptoms among the criteria for ASD.¹⁰ Atypical speech patterns include—but are not limited to—echolalia, pronoun reversal, idiosyncratic and stereotyped speech accompanied by atypical intonation, volume, rhythm, or rate of speech. One of the core features of autism is pragmatic language deficits that are observed across the entire autism spectrum in individuals with varying levels of intellectual and linguistic abilities.¹¹ Pragmatic language refers to the social use of language in everyday interactions and is essential for communicating one's thoughts,

Significant outcomes

- We apply two state-of-the-art machine learning methods to detect autism spectrum disorder (ASD) from two datasets of transcribed statements, collected using two different diagnostic tools: thought, language, and communication (TLC) and autism diagnosis observation schedule-2 (ADOS-2). We report high classification performance on TLC and average performance on ADOS-2, in contrast to previous studies.
- The type of text data is crucial to the effectiveness of models.
- Data augmentation and knowledge exchange between datasets (cross-dataset setting) do not have a positive effect on the results achieved by the models.

Limitations

- In terms of input representation, the limitation of our experiments is the inability to use the structure of diagnostic tools, for example, pagination in ADOS-2 or specific questions of TLC. In other words, our approach results in one aggregated embedding vector reflecting all ADOS-2 pages or all TLC questions, and consequently, question-specific or page-specific signals may be lost.
- The size of our samples, although large by clinical standards, is barely sufficient for machine learning. Deep learning models are typically trained on tens of thousands of observations or more.

ideas, and feelings. It includes using language for various social purposes and the need to produce speech for diverse social contexts. Examples of pragmatic skills include adjusting one's speech to the audience and situation, using humor and irony, sharing information, sustaining conversation, and narrative skills.

Previous works that explored a similar scenario to ours include,¹² where the authors trained models on one dataset and applied it to another. However, the scenario was a cross-disorder one, between schizophrenia (SCZ) and ASD. The results indicated that training the model on the statements of patients with schizophrenia improves the accuracy of recognizing people with ASD, while training it on the statements of people with ASD does not improve the accuracy of recognizing people with SCZ. The paper used an older generation of text encoders

such as the Universal Sentence Encoder¹³ (not trained) with a trainable classification model on top of it.

1.1 | Current study

In this work, we want to use two unique clinical text datasets related to ASD. Each comes from a different diagnostic tool: thought, language, and communication (TLC)¹⁴ and autism diagnosis observation schedule-2 (ADOS-2).¹⁵ Both sets are equal and balanced, which means that they contain the same number of people with the diagnosis as people from the control group. Their unique feature is that each of them allows the use of machine-learning methods. In the case of methods of this type, the often assumed minimum threshold is 50 observations.* Unfortunately, numbers of this order are rarely achieved in clinical settings, where the data collected is often of the order of several or a dozen observations. Unfortunately, this makes it impossible to use machine learning.

Our first goal is to explore the extent to which selected state-of-the-art machine learning and natural language processing tools allow automatic classification of texts as coming from a person with ASD or a control group. On this occasion, we will also examine whether automatic augmentation of the training dataset improves models that recognize the statements of people with ASD.

Our second goal is to test whether knowledge about the characteristics of utterances produced by autistic people can be transferred from one dataset to another. In other words, we will check whether training the model on the TLC set allows us to use this information to improve the classification quality of the model on the ADOS-2 set, and vice versa. Transfer of knowledge between both datasets and diagnostic tools seems possible, especially since both datasets come from people with the same type of disorder. However, this still needs to be investigated. It is also possible that the different types of utterances and characteristics of both ADOS-2 and TLC make such a transfer impossible for the tested machine-learning models.

2 | METHODS

2.1 | Participants

2.1.1 | Sample 1 (TLC)

The first sample comprised 50 participants: 25 adults with a clinical diagnosis of ASD without co-occurring

intellectual disability (according to ICD-10 valid at the time of diagnosis) and 25 demographically matched healthy controls (Table 1). ASD diagnoses were confirmed with the Polish version of the ADOS-2 (Chojnicka and Pisula¹⁵), which was completed by a certified diagnostician. In compliance with ADOS guidelines, in four cases in which assessment was already performed in adulthood, the examination was not repeated and the score was obtained from the diagnostic center via written consent given by the participant. ASD participants were recruited from local therapy centers, support groups, and internet groups. In the case of the control group, participants were recruited from volunteers with no history of neurological or psychiatric disorders who responded to online advertisements. They were paired with patients with ASD based on their sex, age and parental education. Textual utterances in this sample were collected using the TLC Scale.

2.1.2 | Sample 2 (ADOS-2)

The second sample comprised 50 participants: 25 individuals with idiopathic ASD without co-occurring disorder of intellectual development and 25 neurotypical controls. Participants were matched for age, gender, and ethnicity, as well as verbal and nonverbal intelligence quotients (Table 2). Inclusion criteria for participants were as follows: (1) age ≥ 7 years, (2) non-verbal IQ ≥ 80 ; (3) fluency in Polish as a first and primary language; (4) absence of hearing, sight, and mobility impairments; and (5) for the ASD Group, a clinical diagnosis of ASD alongside meeting criteria for autism spectrum on the ADOS-2, Autism Diagnostic Interview-Revised (ADI-R), and Social Communication Questionnaire (SCQ) assessments. Exclusion

TABLE 1 Demographics and clinical characteristics of the Sample 1.

	ASD S1 group (n = 25)	TD group (n = 25)
Females n	16	15
Chronological age in years M (SD)	29.84 (6.39)	30.4 (6.08)
Nonverbal IQ M (SD)	125.44 (10.22)	N/A
ADOS-2 overall total M (SD)	10.67 (5.11)	N/A
Verbal fluency raw score M (SD)	28.44 (7.16)	N/A
Verbal fluency standardized M (SD)	-0.14 (1.16)	N/A

*One example is the well-known scikit-learn cheat sheet. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

TABLE 2 Demographics and clinical characteristics of Sample 2.

	ASD S2 group (<i>n</i> = 25)	TD S2 group (<i>n</i> = 25)	p-Value
Females <i>n</i>	3	3	
Chronological age in years <i>M</i> (<i>SD</i>)	14.55 (5.46)	14.38 (5.83)	0.92
Nonverbal IQ <i>M</i> (<i>SD</i>)	109.08 (13.04)	114.64 (12.50)	0.13
Verbal IQ <i>M</i> (<i>SD</i>)	108.28 (18.30)	113.12 (14.33)	0.20
ADOS-2 overall total <i>M</i> (<i>SD</i>)	13.60 (5.23)	2.48 (3.22)	<0.001
SCQ overall total <i>M</i> (<i>SD</i>)	22.19 (6.01)	3.60 (5.78)	<0.001

criteria for the TD Group were: personal or family history of neurodevelopmental or psychiatric disorders or suspected developmental issues. Textual utterances in this sample were collected using the Book Task from ADOS-2.

2.2 | Measures

2.2.1 | Scale for the assessment of thought, language, and communication

The TLC Scale¹⁴ contains six questions, four of which, respectively, concern the patient and their family, the person closest to the patient, and their interests and childhood. Two of the questions are more abstract—they ask why people get sick and why people believe in God. The responses were recorded and then transcribed. Two variants of the transcriptions were prepared for the patient group—a full version and an edited one (fragments were removed in which patients explicitly admitted to being diagnosed with ASD). The TLC scale was performed by an experimenter at the Institute of Psychology of the Polish Academy of Sciences.

2.2.2 | Autism diagnosis observation schedule, second edition

We also obtained language samples from the ADOS-2 Telling a Story from a Book task (Sample 2). The ADOS-2 is a standardized, semi-structured instrument allowing assessment of social interaction, communication, play/imaginative use of materials, and patterns of behaviors.¹⁶ The Telling a Story from a Book task assesses the participant's ability to narrate a sequential story from a book of pictures. It also provides a context for comments about social relationships, characters' feelings, and responses to conventional humor. In all assessments, we used the picture book "Tuesday" by David Wiesner. The book portrays the adventures of frogs who, one Tuesday evening,

fly to the nearest town on lily pads. The illustrations depict unreal and humorous situations, capturing various mental and emotional states of the characters. According to the procedure, the examiners could give prompts to encourage participants to describe a story, such as "I wonder what happens next." However, examiners were instructed not to label characters' emotions in their prompts.

Recorded narrations were transcribed by two experienced, trained transcribers who were blind to group status and were trained to greater than 80% reliability. One of the transcribers reviewed the transcripts, aligning them with audio recordings to resolve discrepancies.

2.3 | Machine learning for detecting autism

This section describes machine learning models used to detect autism from textual utterances.

2.3.1 | HerBERT

The first approach is based on fine-tuning a pretrained transformer neural network to recognize textual utterances written by people with ASD, where the task is posed as text classification.

Pretraining involves teaching the model a broad understanding of language from huge datasets while fine-tuning adapts this knowledge to specific tasks (also named downstream tasks). The knowledge acquired in pretraining improves the model performance in text classification on downstream task data.

The networks typically used in such scenarios are based on transformer architecture. The basic transformer building block is the self-attention layer,¹⁷ inspired by human cognitive attention. Self-attention utilizes three weight matrices: the query, the values, and the keys, adjusted as model parameters during training. The three matrices can be considered as a single attention head and

are used to transform the input sequence. Multiple stacked transformer layers form an encoder block.

The most well-known application of pretraining to the transformer architecture is BERT,¹⁸ which stands for Bidirectional Encoder Representations from Transformers. The BERT model was pretrained on simple tasks such as predicting a word given its context and predicting the next sentence given the previous one. When first released in 2019, BERT set several records for multiple language-based tasks such as whole text or sequence classification, demonstrating the power and flexibility of such simple pretraining. BERT was originally implemented in the English language in two model sizes: base, of 12 encoders with 12 bidirectional self-attention heads totaling 110M parameters, and large, of 24 encoders with 16 bidirectional self-attention heads totaling 330M parameters.

In our experiments we used HerBERT, which is a monolingual, Polish-only BERT-based Language Model trained using masked language modeling (MLM) and sentence structural objective (SSO) with dynamic masking of whole words.¹⁹ HerBERT was trained on six different corpora available for the Polish language: CCNet Middle and Head, the National Corpus of Polish, Open Subtitles and Wikipedia in Polish, and Wolne Lektury (a collection of school books). The training dataset was tokenized into subwords using a character-level byte-pair encoding with a vocabulary size of 50k tokens. We used two variants: large (allegro/herbert-large-cased) with 330M parameters and base (allegro/herbert-base-cased) with 110M parameters. We used a batch size 4 and tested the learning rates of 1e-4, 1e-5, and 1e-6, a typically recommended range of values. We report the setting of 1e-5, the best performer.

To perform text classification using the HerBERT model, we used the HuggingFace Transformers library,²⁰ which provides tools for common tasks such as text classification. The implementation aimed at text classification (class is named `BertForSequenceClassification`), is applied on top of the final hidden state of the [CLS] token, and contains a dropout layer followed by a linear layer. This approach allows for training the whole model for the task at hand (110M or 330M parameters depending on the HerBERT variant), not only the classification head. Figure 1 shows a simplified diagram of the HerBERT model architecture, composed of 12 encoder layers for the base variant and 24 layers for the large variant, followed by a linear classifier layer.

2.3.2 | ada + SVM

In the second approach, we used two separate steps.

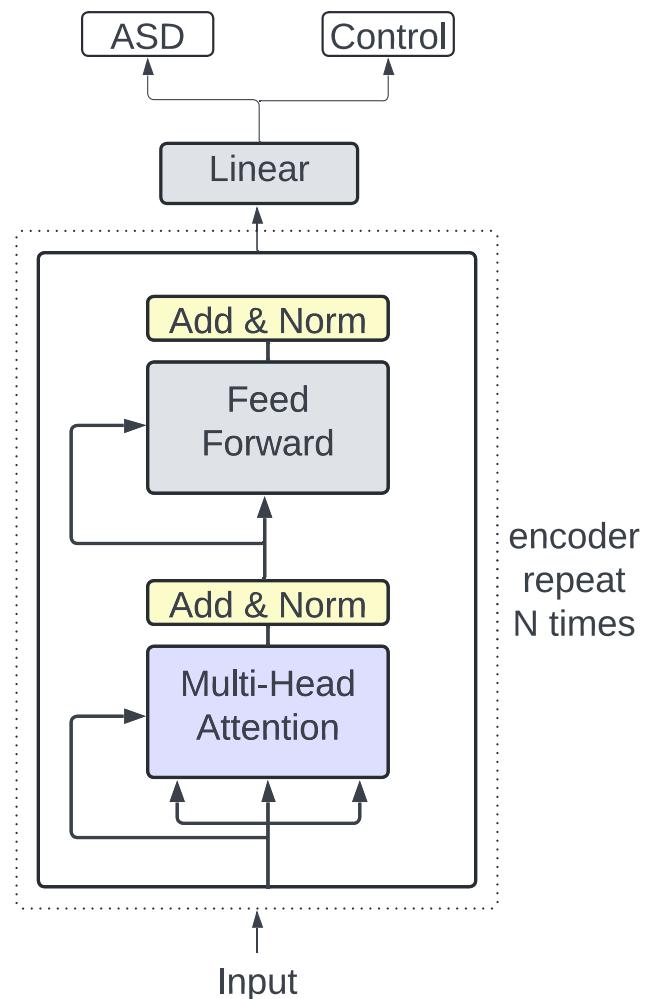


FIGURE 1 A simplified diagram of a HerBERT-based classifier.

In the first step, we computed the embedding vector representation of each of the texts in our datasets. Embedding vectors (embeddings) are numerical representations of concepts; they reflect texts converted to sequences of floating point numbers. A popular approach to compute embeddings is to use pretrained transformer neural networks. Specifically, embeddings can be extracted from the last layer of BERT-style networks. In most cases, embeddings originate from models that were just pretrained and not fine-tuned to any downstream task. Their goal is to be usable in many scenarios, such as text classification, question answering, and information retrieval.

In the second step, we train a classifier model on such broadly usable embeddings. Such a classifier is lightweight and consists of thousands of trainable parameter weights, as opposed to hundreds of millions as in the BERT model. For our experiments, we selected the leading multilingual embeddings, namely OpenAI's text-ada-002, which can embed up to approximately 6000 words into a 1536-dimensional vector. While powerful, text-ada-

TABLE 3 TLC results: accuracy \pm standard error of cross-validation measure, sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). All reported model results are averaged over 10-fold cross-validation. Data augmentation variants are denoted as aug.

Model	Accuracy \pm std err	Sensitivity	Specificity	PPV	NPV
HerBERT-base	0.76 \pm 0.06	0.72	0.80	0.78	0.74
HerBERT-base + aug	0.68 \pm 0.07	0.60	0.76	0.71	0.66
HerBERT-large	0.68 \pm 0.07	0.44	0.92	0.85	0.62
HerBERT-large + aug	0.72 \pm 0.06	0.72	0.72	0.72	0.72
ada + SVM	0.74 \pm 0.06	0.72	0.76	0.75	0.73

TABLE 4 ADOS-2 Book Task results: accuracy \pm standard error of cross-validation measure, sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). All reported model results are averaged over 10-fold cross-validation. Data augmentation variants are denoted as aug.

Model	Accuracy \pm std err	Sensitivity	Specificity	PPV	NPV
HerBERT-base	0.52 \pm 0.07	0.44	0.60	0.52	0.52
HerBERT-base + aug	0.48 \pm 0.07	0.48	0.48	0.48	0.48
HerBERT-large	0.58 \pm 0.07	0.52	0.64	0.59	0.57
HerBERT-large + aug	0.54 \pm 0.07	0.48	0.60	0.55	0.54
ada + SVM	0.42 \pm 0.07	0.40	0.44	0.42	0.42

002 is not open source and is only available via API.[†] Thus, users must have an Internet connection to query their text-ada-002 databases. Additionally, this introduces API costs and users are locked into one vendor. Unfortunately, no research paper on OpenAI embeddings is available and we can only speculate that the solution might be based on extracting embeddings from a transformer neural network. For classification, we selected the well-known support vector machine (SVM) algorithm with a radial kernel.²¹ The choice of SVM was due to the algorithm's robustness and well-proven ability to classify text embedding vectors pre-computed by a neural network-based text encoder. In particular, it proved to be the most accurate of the tested methods in both ASD and SCZ text classification in Reference 12, where it was applied to embeddings computed with the Universal Sentence Encoder.

2.3.3 | Data augmentation

Data augmentation techniques might improve model performance in low-data scenarios by generating additional, synthetic data using the existing dataset. Augmentation methods are popular in computer vision applications, but such techniques can be also used for text processing.

We tested a technique based on back-translation. In this method, we translate each text to some language

(in this case English) and then translate it back to Polish, the original language. This can help generate text data containing different words while maintaining the meaning of the text. In other words, for each utterance the dataset is enriched with its second variant, which is semantically close or identical, but different at the lexical and syntactic levels. The Google Translate API was used to translate the datasets.

3 | RESULTS

3.1 | In-dataset

In the in-dataset scenario, we report the performance of models evaluated in a 10-fold cross-validation using only one dataset for both model training and evaluation. For HerBERT model validation, at each fold, the validation set consisted of three randomly sampled observations (texts) taken from the training part of the set, which is equivalent to 5% of the data. To select the best model, we used the accuracy reported on this validation set. Table 3 contains the results of TLC and Table 4 the results of ADOS-2 in-domain experiments.

3.2 | Cross-dataset

In the cross-dataset scenario, we tested knowledge transfer between both datasets—the results are available in

[†]Accessed on 1.03.2024 <https://platform.openai.com/docs/guides/embeddings>

TABLE 5 Cross-dataset results: We trained on one (source) dataset and tested on another (target) dataset, with optional training on the target dataset.

Model (learning rate)	Variant	Accuracy \pm std err	Sensitivity	Specificity	PPV	NPV
TLC to ADOS						
HerBERT-base (1e-5)	Full	0.42 \pm 0.07	0.16	0.68	0.33	0.45
HerBERT-base (1e-6)	Full	0.46 \pm 0.07	0.56	0.36	0.47	0.45
HerBERT-base (1e-5)	Light	0.54 \pm 0.07	0.24	0.84	0.6	0.53
HerBERT-base (1e-6)	Light	0.46 \pm 0.07	0.8	0.12	0.48	0.38
HerBERT-base (1e-5)	0-shot	0.42	0.64	0.2	0.44	0.36
ada + SVM	0-shot	0.50 \pm 0.07	1	0	0.5	0
ADOS to TLC						
HerBERT-large (1e-5)	Full	0.66 \pm 0.07	0.56	0.76	0.7	0.63
HerBERT-large (1e-6)	Full	0.60 \pm 0.07	0.48	0.72	0.63	0.58
HerBERT-large (1e-6)	Light	0.50 \pm 0.07	0	1	0	0.5
HerBERT-large (1e-5)	Light	0.50 \pm 0.07	0	1	0	0.5
HerBERT-large (1e-5)	0-shot	0.5	0	1	0	0.5
ada + SVM	0-shot	0.48 \pm 0.07	0.84	0.12	0.49	0.43

Note: The table reports the following metrics: accuracy \pm standard error of cross-validation, sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). In the variant 0-shot, training is performed only on the source dataset, models are not trained on the target data. The reported results come from one model. In variants full and light, models are initially trained on the source dataset, and then the training (and testing) is continued on the target dataset. Results are then averaged over 10-fold cross-validation on the target data. In the full variant, training on the target dataset is 10 epochs, and the variant light is just 1 epoch.

Table 5. First, we examined zero-shot model performance. Zero-shot text classification refers to a task where a model is trained on a set of labeled examples of one type (the first dataset) but is then used to classify new examples from previously unseen text types (the second dataset). In the zero-shot setting, there is no model training on the second dataset.

Second, we examined the performance of models in a scenario where one dataset is used for pretraining and another one is used for fine-tuning (typical model training). As in the in-domain scenario, we report the performance of models evaluated on the target dataset in a 10-fold cross-validation. For HerBERT model validation, at each fold, the validation set consisted of three randomly sampled observations (texts) taken from the training part of the set. Again, to select the best model, we used the accuracy reported on this validation set. We tested the training process in two variants, called full and light, where the training is either 10 epochs or one epoch long. One epoch-long (light) training should be less prone to forgetting the previously learned knowledge. We also test two learning rates: 1e-5 (the best one for regular model training) and 1e-6 (a low one to avoid potential forgetting of the information acquired during pretraining). In the case of HerBERT models, we selected the variant that performs the best on the

source dataset. For example, we selected the HerBERT-large to try on TLC data, as it was the one that performed best on ADOS-2.

4 | DISCUSSION

Prior research indicates that narrative difficulties comprise a core component of social communication deficits in autism, regardless of heterogeneity in the language capacities among individuals with ASD.²² In studies on narrative abilities in autism, various narrative tasks have been used, including personal narratives, story recall, and most commonly, storytelling. We investigated the role of narrative context in the analysis involving autism detection using machine learning based on textual data. We compared two samples, for which speech data was collected in two different narrative contexts. One group responded to six open-ended questions comprising the TLC Scale. The other group narrated a story depicted in pictures in a book task from the ADOS-2 assessment.

Our paper demonstrated the good performance of automated ASD detection when applied to TLC texts. The best results were obtained by the smaller HerBERT variant, HerBERT-base. The accuracy was as high as 0.76, with reasonably high values of the other four metrics. This was the single best result of all experiments

described in this paper. Notably, the ada+SVM technique turned out to be not far behind, reaching an accuracy of 0.74. Yet, the comparison of both methods is difficult since the standard error of cross-validation was at 6–7 percentage points.

Interestingly, for the ADOS-2 data the best variant was HerBERT-large. The performance was not very high, with the accuracy reaching 0.56. Future study needs to investigate why the previous generation of pretrained models described in Reference 23 achieved better performance on this dataset. Possible explanations could include not utilizing the data structure, as was the case in Reference 23: we "compressed" the representation of each person's entire utterance into a single embedding (as this was the only option to perform cross-dataset experiments), yet the most effective method in Reference 23 was to individually embed dialogs linked to each page of the picture book. The performance of "compressed," whole utterance embedding reached a similarly bad performance as in our experiments.

The effect of data augmentation was consistently bad, decreasing the performance of each tested model configuration. It may indicate that autism identification is likely based on subtle clues that are lost in translation; machine translating ignores such clues and reflects the most common, typical language patterns.

The cross-dataset scenario did not yield the expected improvements. Generic models initialized from HerBERT checkpoints (without pretraining on another ASD dataset) performed better than models that were pretrained on another ASD dataset. This could mean that the differences between both datasets (in terms of information relevant to detecting ASD utterances) are more pronounced than the similarities. It is also possible that the cross-dataset knowledge transfer failed due to too coarse-grained embeddings, which might be the issue for ADOS-2.²³

The BERT model had significantly higher accuracy in classifying people from Sample 1 (TLC) than 2 (ADOS-2). One of the differences between these groups is age—people from the second sample are younger. Williams et al.²⁴ compared brain function in children and adults with autism in a task requiring language processing. The children and adults with autism differed from each other in the use of some brain regions during the task, but the adults with autism had activation levels similar to those of the control groups. According to the authors, the differences between the two autism age groups may be indicative of positive changes in neural function related to language processing associated with maturation and/or educational experience. Although findings from an MRI study on language processing do not directly allow for conclusions about the language differences

between autistic adults and children, it seems that the language of autistic adults may be more similar to that of neurotypical people than that of autistic children, which should make the automated classification of older participants more difficult. Other studies have also indicated an increase in the language skills of autistic participants with age. For instance, interesting findings on changes in cognitive and language skills during development were provided by the 40-year follow-up by Howlin et al.²⁵ The authors assessed 60 autistic individuals with an IQ in the average range as children. Language abilities improved from childhood to adulthood. However, the authors did not use strictly linguistic tests but rather scores derived from the language and communication scale of the ADI-R. The study by McIntyre et al.²⁶ on changes in narrative skills in autistic individuals aged 8 to 16 years old without disorders of intellectual development also indicated an improvement with age. However, as the authors pointed out, age was not a significant covariate in performed analyses after controlling for ASD symptom severity and lexical-semantic knowledge.

Interestingly, in the first sample there were proportionally more women, both on the autism spectrum and neurotypical, than in the second sample. Differences in the speech of women and men have long been recognized, including those detected by automated computational methods. Although depending on the analyzed aspects of speech and language as well as the methodology, the research results vary (e.g., References 27,28). An increasing number of studies point to differences between the narrative skills of autistic boys and girls.^{29,30} Boorse et al.²⁹ showed a unique narrative profile of autistic girls that overlapped with autistic boys and typical girls/boys concerning the number of nouns and cognitive process words (e.g., "think," "know") in their stories. Participants were 7–15 years old, verbally fluent, and without disorders of intellectual development and produced narrations during the ADOS-2 Cartoons task. The authors reported that autistic children of both sexes used more nouns in their narrations than their neurotypical peers, indicating object-focused storytelling. However, autistic girls differed from autistic boys by producing a greater number of cognitive process words. The recent study by Cola et al.³¹ indicates that autistic girls use more social words and particularly more friend words than autistic boys during interviews conducted as part of the ADOS-2 assessment. Social words were defined as words that make reference to other people (e.g., "classmates," "everyone," or "them"), whereas friend words were defined as words that refer to both friends and peers (e.g., "buddies," "best friend"). Participants were 6–15 years old, verbally fluent, and without disorders of intellectual development. In our study, the

group sizes did not allow us to examine sex differences; however, the larger number of women in Sample 1 where the TLC interview was used did not decrease the effectiveness of automated detection.

Therefore, it appears that the differences in model performance might be related to narrative context. The Book task from the ADOS-2 assesses an individual's ability to tell a sequential story from a book of pictures.¹⁶ Thus the produced narration is fictional, as opposed to personal narration in the TLC interview. The nature of the Book task means that some words in the utterances of autistic participants and controls will be the same, namely the characters or elements visible in the pictures. As a consequence, this may reduce the differences between groups compared with a less structured TLC interview. The TLC interview, where open-ended questions are not supplemented by visual stimuli, gave the respondents much greater freedom and guaranteed a greater variety of statements. Prior evidence suggests the role of narrative context: autistic participants exhibited more difficulties in less structured tasks, such as narrating personal experiences or describing pictures from the Thematic Apperception Test, compared with picture book tasks.^{32,33} However, these observations were not confirmed by a meta-analysis conducted by Baixauli et al.,²² who did not find statistically significant differences depending on the type of narrative (picture book/fictional storytelling vs. autobiographical/personal/everyday activities stories).

Model performance for Sample 2 (ADOS-2) is much lower than the sensitivity and specificity values that we obtained for the same narrative task in our previous study.²³ We examined two text encoders: embeddings from language models (ELMo)³⁴ and Universal Sentence Encoder (USE),³⁵ and three classification algorithms: XG Boost,³⁶ SVMs,³⁷ and Dense neural network layer.³⁸ Both encoders, ELMo and USE, achieved sensitivities exceeding 0.70 and a specificity of 0.68. However, the setup of the previous study is not entirely comparable. In Reference 23, we tested a setting in which we used the ADOS page structure: for each of the 16 pages, we computed a separate embedding vector. In this way, embeddings were sensitive to specific signals appearing in the context of a specific page. The variant with a single text, which is a concatenation of all pages and one embedding vector, turned out to be worse than the variant with many pages and vectors. For example, the accuracy score achieved by the USE model with the SVM classifier in the single vector setting was only 0.58. However, in the current study, we cannot use multi-page splitting and separate vectors because we want the vectors to be universally applicable to both ADOS-2 and TLC. For this purpose, concatenation and one embedding vector representing the entire statement are necessary.

Researchers working on clinical samples in the field of machine learning face the challenge of relatively small datasets—significantly smaller than those typically employed in other types of AI-driven analyses. There have been attempts to establish criteria for evaluating the study sample size in machine learning (e.g., Reference 39). Consequently, scientists sometimes opt to employ data augmentation to increase the quantity of data supplied to neural networks. We decided to compare the effectiveness of detection also for data subjected to augmentation. We chose the back translation approach, wherein the text is machine-translated into another language (English in our case) and then back into the original language (Polish in our case). This approach allows for a doubling of the number of observations. However, caution is strongly advised, as different psychiatric conditions are associated with certain linguistic features that may be lost during the process. The results we obtained confirmed that this may indeed apply to the autism spectrum. Detection in the case of augmented data was significantly less effective than for the original utterances, regardless of narrative context. Various atypical features can be observed in the speech of autistic individuals, which may be lost in the process of translation and back-translation. Among these, idiosyncratic or stereotypical language, non-obvious word combinations, or the use of words with lower frequency in the language can be mentioned.⁹ Machine translation carries the risk of “polishing” the text to obtain the most typical and linguistically correct formulations. Also, the language of autistic individuals may exhibit a pedantic quality or be described as overly formal and adult-like.⁴⁰ We argue that augmentation of textual data in the case of autism spectrum may result in the loss of valuable linguistic characteristics.

The results of this study highlight the problem of model performance when transferring to another dataset, along with the lack of insight into the reasons behind decisions made by a model. This is a significant challenge for researchers, hindering the clinical applicability of the tested models. In addition to the issues described above—such as small clinical sample sizes relative to the needs of machine learning models and varying performance levels among demographic groups—another problem involves distinguishing autism from related conditions. The vast majority of previous studies compared autistic participants to neurotypical ones. A challenge for future research is to attempt to distinguish between various psychiatric conditions with overlapping symptoms, which will be much more similar to the challenges clinicians face during the diagnostic processes.

The obtained results are promising for the application of automated methods in studies on the autism spectrum. Although further research is needed before such models

can be implemented in clinical settings, one can already envision their use for screening purposes or as tools to support clinicians in decision-making processes, akin to today's standardized paper-and-pencil psychometric tools. With the advancement of machine learning methods, automated autism classification is bound to improve and become more widespread.

To conclude, we applied two state-of-the-art machine learning methods—a monolingual Polish-language BERT variant (HerBERT) and Ada OpenAI embeddings followed by an SVM classifier—to detect ASD from datasets of transcribed statements collected using two diagnostic tools: TLC and ADOS-2. We obtained high classification performance on TLC and average performance on ADOS-2 data, in contrast to previous studies.²³

We experimented with a cross-dataset setup, where we predicted ASD on one dataset while training models on another. The lack of promising results indicates that the clues used by the models are dataset-specific and non-transferable to other scenarios, quite likely due to the different structure of each of the two datasets.

The implications of our study are as follows.

We confirmed that automated detection of ASD from textual utterances is a promising direction. However, real-world usage of machine-learning models based on the current generation of pretrained neural networks, despite their promising results, needs further research on explaining model decisions to human practitioners.

We also point to the fact that to obtain top ASD detection performance, it is advisable to focus on one data type when training models, explore its dataset-specific structure to use more fine-grained data representations, of questions (TLC) or utterances linked to page numbers (ADOS-2), as opposed to aggregated embeddings of whole utterances.

FUNDING INFORMATION

The project was supported by grants from the National Science Centre (Poland) (grant no. 2020/39/D/HS6/00809 awarded to Izabela Chojnicka and grant no. 2018/31/B/HS6/02848 awarded to Małgorzata Krawczyk) and also from the funds awarded by the Ministry of Science and Higher Education (Poland) in the form of a subsidy for the maintenance and development of research potential in 2023 (501-D125-01-1250000 zlec.5011000231; 5011000228). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets are available from the authors upon request. All source code (python) files will be uploaded to Github upon the acceptance of our paper.

ETHICS STATEMENT

This study was conducted by the Declaration of Helsinki, and approved by the Ethics Committee of the Institute of Psychology, Polish Academy of Sciences (decision number: 17/X/2019) for studies involving humans.

PATIENT CONSENT STATEMENT

Informed consent was obtained from all subjects involved in the study.

ORCID

Aleksander Wawer  <https://orcid.org/0000-0002-7081-9797>

REFERENCES

1. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2020. *MMWR Surveill Summ.* 2023;72(2):1-14.
2. Baio J, Wiggins L, Christensen DL, et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveill Summ.* 2018;67(6):1-23.
3. Rice C. Prevalence of autism spectrum disorders—autism and developmental disabilities monitoring network. PsycEXTRA Dataset 2009. 2006.
4. Washington P, Wall DP. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. *Ann Rev Biomed Data Sci.* 2023;6(1):211-228.
5. Kim JH, Hong J, Choi H, et al. Development of deep ensembles to screen for autism and symptom severity using retinal photographs. *JAMA Netw Open.* 2023;6(12):e2347692. doi:[10.1001/jamanetworkopen.2023.47692](https://doi.org/10.1001/jamanetworkopen.2023.47692)
6. Shrivastava T, Singh V, Agrawal A. Autism spectrum disorder detection with kNN imputer and machine learning classifiers via questionnaire mode of screening. *Health Inf Sci Syst.* 2024; 12(1):18. doi:[10.1007/s13755-024-00277-8](https://doi.org/10.1007/s13755-024-00277-8)
7. Nabil MA, Akram A, Fathalla KM. Applying machine learning on home videos for remote autism diagnosis: further study and analysis. *Health Inform J.* 2021;27(1):146045822199188.
8. Wei Q, Cao H, Shi Y, Xu X, Li T. Machine learning based on eye-tracking data to identify autism spectrum disorder: a systematic review and meta-analysis. *J Biomed Inform.* 2023;137:104254.
9. Vogindroukas I, Stankova M, Chelas EN, Proedrou A. Language and speech characteristics in autism. *Neuropsychiatr Dis Treat.* 2022;18:2367-2377.
10. World Health Organization. *International statistical classification of diseases and related health problems.* World Health Organization; 2021 Accessed February 22, 2024. <https://icd.who.int/browse11/l-m/>
11. Tager-Flusberg H. Defining language phenotypes in autism. *Clin Neurosci Res.* 2006;6(3-4):219-224.
12. Wawer A, Chojnicka I, Okruszek L, Sarzynska-Wawer J. Single and cross-disorder detection for autism and schizophrenia. *Cogn Comput.* 2022;14:1-13.
13. Yang Y, Cer D, Ahmad A, et al. Multilingual universal sentence encoder for semantic retrieval. In: Celikyilmaz A, Wen TH, eds. *Proceedings of the 58th Annual Meeting of*

- the Association for Computational Linguistics: System Demonstrations Online.* Association for Computational Linguistics; 2020:87-94 <https://aclanthology.org/2020.acl-demos.12>
14. Czernikiewicz A. Przewodnik po zaburzeniach językowych w schizofrenii. *Instytut Psychiatrii i Neurologii*; 2004.
 15. Chojnicka I, Pisula E. Adaptation and validation of the ADOS-2, polish version. *Front Psychol.* 2017;8:1916.
 16. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop SL. *Autism Diagnostic Observation Schedule*. WPS; 2012.
 17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates; 2017 https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf
 18. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol 1. Association for Computational Linguistics; 2019:4171-4186 <https://aclanthology.org/N19-1423>
 19. Mroczkowski R, Rybak P, Wróblewska A, Gawlik I. HerBERT: efficiently pretrained transformer-based language model for polish. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing Kiyv*. Association for Computational Linguistics; 2021:1-10 <https://www.aclweb.org/anthology/2021.bsnlp-1.1>
 20. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations: Association for Computational Linguistics*. 2020;38-45. <https://aclanthology.org/2020.emnlp-demos.6/>
 21. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273-297.
 22. Baixaulli I, Colomer C, Roselló B, Miranda A. Narratives of children with high-functioning autism spectrum disorder: a meta-analysis. *Res Dev Disabil.* 2016;59:234-254.
 23. Wawer A, Chojnicka I. Detecting autism from picture book narratives using deep neural utterance embeddings. *Int J Lang Commun Disord.* 2022;57(5):948-962. doi:[10.1111/1460-6984.12731](https://doi.org/10.1111/1460-6984.12731)
 24. Williams DL, Cherkassky VL, Mason RA, Keller TA, Minshew NJ, Just MA. Brain function differences in language processing in children and adults with autism. *Autism Res.* 2013;6(4):288-302.
 25. Howlin P, Savage S, Moss P, Tempier A, Rutter M. Cognitive and language skills in adults with autism: a 40-year follow-up. *J Child Psychol Psychiatry.* 2014;55(1):49-58.
 26. McIntyre NS, Grimm RP, Solari EJ, Zajic MC, Mundy PC. Growth in narrative retelling and inference abilities and relations with reading comprehension in children and adolescents with autism spectrum disorder. *Autism Dev Lang Impair.* 2020; 5:2396941520968028.
 27. Newman ML, Groom CJ, Handelman LD, Pennebaker JW. Gender differences in language use: an analysis of 14,000 text samples. *Discourse Process.* 2008;45(3):211-236.
 28. Piersoul J, Van de Velde F. Men use more complex language than women, but the difference has decreased over time: a study on 120 years of written Dutch. *Linguistics.* 2022;61(3): 725-747.
 29. Boorse J, Cola M, Plate S, et al. Linguistic markers of autism in girls: evidence of a “blended phenotype” during storytelling. *Mol Autism.* 2019;10(1):1-12.
 30. Kauschke C, van der Beek B, Kamp-Becker I. Narratives of girls and boys with autism spectrum disorders: gender differences in narrative competence and internal state language. *J Autism Dev Disord.* 2015;46(3):840-852.
 31. Cola M, Yankowitz LD, Tena K, et al. Friend matters: sex differences in social language during autism diagnostic interviews. *Mol Autism.* 2022;13:1-16.
 32. Lee M, Nayar K, Maltman N, et al. Understanding social communication differences in autism spectrum disorder and first-degree relatives: a study of looking and speaking. *J Autism Dev Disord.* 2020;50:2128-2141.
 33. Losh M, Capps L. Narrative ability in high-functioning children with autism or Asperger's syndrome. *J Autism Dev Disord.* 2003;33:239-251.
 34. Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol 1. Association for Computational Linguistics; 2018:2227-2237 <http://aclweb.org/anthology/N18-1202>
 35. Cer D, Yang Y, Sy K, et al. *Universal Sentence Encoder for English*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations Brussels, Belgium: Association for Computational Linguistics; 2018:169-174 <https://www.aclweb.org/anthology/D18-2029>
 36. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'16*. ACM; 2016: 785-794.
 37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
 38. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015 <https://www.tensorflow.org/>
 39. Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics.* 2023;24(1):48.
 40. Luyster RJ, Zane E, Wisman WL. Conventions for unconventional language: revisiting a framework for spoken language features in autism. *Autism Dev Lang Impair.* 2022;7:1-19.

How to cite this article: Wawer A, Chojnicka I, Sarzyńska-Wawer J, Krawczyk M. A cross-dataset study on automatic detection of autism spectrum disorder from text data. *Acta Psychiatr Scand.* 2024; 1-11. doi:[10.1111/acps.13737](https://doi.org/10.1111/acps.13737)