



RESEARCH REPORT

Detecting autism from picture book narratives using deep neural utterance embeddings

Aleksander Wawer^{1,*}  | Izabela Chojnicka^{2,*} 

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

²Faculty of Psychology, University of Warsaw, Warsaw, Poland
(Email: izabela.chojnicka@psych.uw.edu.pl)

Correspondence

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

*Equal Contribution

Abstract

Background: Deficits in the ability to use language in social contexts, including storytelling skills, are observed across the autism spectrum. Development in machine-learning approaches may contribute to clinical psychology and psychiatry, given its potential to support decisions concerning the diagnosis and treatment of psychiatric conditions and disorders.

Aims: To evaluate the usefulness of deep neural networks for detecting autism spectrum disorder (ASD) from textual utterances, specifically from narrations produced by individuals with ASD.

Methods & Procedures: We examined two text encoders: Embeddings from Language Models (ELMo) and Universal Sentence Encoder (USE), and three classification algorithms: XGBoost, support vector machines, and dense neural network layer. We aimed to classify 25 participants with ASD and 25 participants with typical development (TD) based on their narrations produced during the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) picture book task. The results of computational approaches were compared with the results of standardized testing and classifications made by two psychiatrists (raters). The raters were asked to read utterances produced by a participant (without an examiner's statements and additional information) and assign a participant to one of the two groups: ASD or with typical development (TD).

Outcomes & Results: The computer-based models had higher sensitivity, specificity, positive predictive values and negative predictive values than the raters, and lower than the two standardized instruments: ADOS-2 and Social Communication Questionnaire (SCQ).

Conclusions & Implications: Our findings lay the groundwork for future studies involving deep neural network-based text representation models as tools for augmenting the ASD diagnosis or screening. Both ELMo and USE text encoders provided promising specificities, sensitivities, positive predictive values and

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *International Journal of Language & Communication Disorders* published by John Wiley & Sons Ltd on behalf of Royal College of Speech and Language Therapists

negative predictive values. Our results indicate the usefulness of page-level embeddings for utterance representation in ADOS-2 picture book task.

KEYWORDS

autism spectrum disorders (ASD), deep neural networks, deep learning, language processing, narrative

WHAT THIS PAPER ADDS

What is already known on this subject

Deficits in the use of language in social contexts, and narrative ability in particular, are observed across the autism spectrum. Most research on narrative skills has applied hand-coding methods. Hitherto, machine-learning methods were used mostly for image recognition problems and data from screening questionnaires for ASD classification. Detection of mental and developmental disorders from textual input is an emerging field for machine and deep-learning methods.

What this paper adds to existing knowledge

This study explored the ability of several types of deep neural network-based text representation models to detect ASD. Both ELMo and USE provided the most promising values of specificity, sensitivity, positive predictive values and negative predictive values.

What are the potential or actual clinical implications of this work?

Competitive accuracy, repeatability, speed and ease of operation are all advantages of computerized methods. They allow for objective and quantitative assessment of narrative ability and complex language skills. Deep neural network-based text representation models could in the future support clinicians and augment the decision-making process related to ASD diagnosis, screening and intervention planning.

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that affects communication, social interactions and is associated with restricted, repetitive patterns of behaviour and interests (World Health Organization (WHO), 2021). The clinical expression of the disorder is heterogeneous, and symptoms may vary widely from person to person. The cognitive abilities of people with ASD can range from gifted to severely challenged (Goldstein & Ozonoff, 2018). The worldwide prevalence of ASD is estimated at 1% (Lai et al., 2014). Since 2000, the prevalence of ASD has increased by approximately 176%, making it an urgent public health concern (Sampaio et al., 2021; Wallace et al., 2012). ASD symptoms last throughout a person's life and affect personal, family, educational and professional experiences. People with ASD may communicate, interact, behave and learn differently from most other people. Many of them need support in daily activities (Bal et al., 2015).

Although contributing genetic variation can be identified in 5–30% of individuals with ASD (Schaaf et al., 2020), ASD is still defined behaviourally and diagnostic assessment is based on information about the signs and patterns of behaviour (WHO, 2021). It is a complex process that takes two steps: screening and comprehensive diagnostic evaluation. Diagnostic assessment is recommended to be multidisciplinary and should be based on information from many sources, such as a developmental interview with a parent or caregiver, information from community settings (e.g., kindergarten), observational assessments, standardized testing and a medical examination (Goldstein & Ozonoff, 2018). The validity and reliability of diagnosis of ASD improves considerably when diagnostic evaluation is carried out using standardized tools with good psychometric properties (Kim & Lord, 2012). The development of computational models, specifically natural language-processing models as the subject of the present study, may provide tools that, similarly to

psychological testing, support clinicians in their decision-making. The advantage of computerized methods includes competitive accuracy, repeatability, speed and ease of operation. However, due to the multidimensionality and complexity of the diagnosis problem, the role of such automatic analyses of narratives may only be complementary or supportive.

Narrative ability in ASD

Since deficits in pragmatic language (the ability to use language in social contexts) are observed across the autism spectrum (Parsons et al., 2017), they became the focus of research in computational linguistics (Chojnicka & Wawer, 2020; Lee et al., 2018; Losh & Gordon, 2014). Among social communication deficits observed in ASD are poor narrative (i.e., storytelling) skills (Baixauli et al., 2016) that are reported universally across languages with different typologies (Engberg-Pedersen & Christensen, 2016; Sah & Torng, 2015; Mäkinen et al., 2014). Individuals with ASD struggle with conveying the gist of the story and use referential devices inappropriately. They produce stories lacking coherence and causal connections with the inclusion of inappropriate or irrelevant components at the same time (Diehl et al., 2006). Structural components of narration are also affected in ASD: narratives of individuals with ASD are characterized by fewer words and utterances and reduced lexical diversity compared with typically developing (TD) peers (Capps et al., 2000).

Another domain of narration studied in ASD, comprising the vocabulary used to describe a character's perceptions, emotions and cognitive states, is the internal state language (ISL). Research indicates that children and adolescents with ASD use fewer ISL terms in their narrations than their peers with TD. Baixauli et al. (2016) described the moderating effect of verbal IQ: higher IQ levels are associated with greater difficulties with ISL. Subclinical narrative difficulties are also seen in parents of individuals with ASD and are considered a part of the broad ASD phenotype (Losh et al., 2008). Thus, pragmatic deficits are hypothesized to constitute a genetically meaningful trait (Losh et al., 2012).

Studies of narrative abilities in ASD typically involve school children, adolescents and adults. Therefore, in many cases, narrative research will not translate directly into diagnostic practice. However, even though ASD can be diagnosed as early as 2 years of age or less, most children are not diagnosed with ASD until after their fourth birthday, and for ASD with no intellectual disability even later (Maenner et al., 2020). There are children with less severe levels of autism spectrum-related symptoms and

no intellectual disability for whom the behavioural signs of autism spectrum might not be clear enough until they are older and have to respond to the social and educational challenges of school and friendships. In addition to the diagnostic process, another type of assessment that occurs during school age is the evaluation of child skills that support a wide range of interventions (Goldstein & Ozonoff, 2018). Moreover, objective, valid and reliable measures of narration ability among individuals with ASD with no clear structural language deficits are limited. Standardized tests often miss evident-in-naturalistic-settings differences (Lee et al., 2018). Considering that narrative skills are used daily in everyday interactions and experiences, accurate assessment of narrative ability may be essential both for the diagnostic process and intervention planning. As Losh and Gordon (2014) argue, the computational approach described in the following subsection may provide a way to implement more diversified narrative tasks in clinical assessments.

Computational models for studying narration

To date, most research on narrative skills has applied hand-coding methods. However, the manual coding of results has considerable limitations: it is time and labour intensive, and requires extensive training for coders to achieve reliability. These drawbacks mean that hand-coding approaches limit sample sizes and may be difficult to implement across different research sites. Only a few papers describe computational approaches to characterize narrative performance in ASD. Lee et al. (2018) compared narrations of 19 individuals with ASD and 14 TD controls using Latent Semantic Analysis (LSA)—a quantitative, natural language processing (NLP) method that determines the semantic similarities of words and bodies of text. Similarly, Losh and Gordon (2014) analysed narrations from 22 individuals with ASD with no intellectual disability and 26 typically developing peers using LSA. Although LSA successfully differentiated participants with ASD from TD controls (Lee et al., 2018; Losh & Gordon, 2014), the certain narrative qualities underlying the results (such as number of words, complex syntax, use of evaluative devices, etc.) were mostly (except semantic similarity) unknown. Therefore, the authors carried out hand-coding analyses of key indices of narrative quality such as total story clauses, complex syntax and evaluative devices, for example, referring to characters' internal states or causal language to bind story events (Lee et al., 2018; Losh & Gordon, 2014).

Chojnicka and Wawer (2020) applied two NLP methods: sentiment and language abstraction analyses to compare the picture book narrations of participants with ASD and with TD. They used the Linguistic Category Model (LCM): an automated, dictionary-based tool to measure language abstraction based on categorizing verbs according to their level of abstraction (Wawer & Sarzyńska, 2018). The second method used was sentiment analysis: a computational technique that extracts subjective information from text in terms of the positive or negative emotional tones of an utterance (Mäntylä et al., 2018). Stories of individuals with ASD were characterized by a lower level of language abstraction than those of participants with TD (Chojnicka & Wawer, 2020). The level of language abstraction was strongly, significantly and positively correlated with words with emotional polarity. Participants with ASD used fewer words with positive sentiment with large effect size and marginally fewer words with negative sentiment.

Deep neural networks

All the above-described studies applied computational techniques other than neural networks in narrative research. In recent years, the interest in using machine-learning algorithms in biomedical fields, including psychiatry, has increased significantly. Among them, deep-learning techniques based on artificial neural networks (ANNs) have been successfully employed in disease diagnosis involving image recognition problems (neuroimaging). Initially, ANN were used to model specific cognitive problems in ASD such as attentional impairments (Gustafsson & Paplinski, 2004), functional disconnections (Park et al., 2019), or alterations of the precision of predictions and sensory information processing (Philippsen & Nagai, 2018). The overview of ANN models used for ASD is presented by Lanillos et al. (2020).

Many examples of successful applications of deep neural networks have been published in recent years, demonstrating their ability to encode both semantics and syntactic properties of input texts. Multilayer neural networks used for this purpose are initially trained on a simple task of language modelling (predicting words from their contexts) to gain knowledge about language. This first phase, known as pre-training and transfer learning, is performed on huge amounts of textual data, typically hundreds of gigabytes and more. After pre-training, vector representation of any input text may be extracted from this network (Cer et al., 2018; Peters et al., 2018). It retains relevant information, which makes it usable for other tasks. Such representation vectors, called sentence embeddings or text embeddings,

are usable for supervised learning as an input for classification algorithms. This second phase is known as fine-tuning (Devlin et al., 2018). Thanks to the properties of embedding vectors, classification algorithms can successfully learn from relatively small labelled data sets. Examples of neural networks suitable for generating accurate utterance embeddings include Embeddings from Language Models (ELMo; Peters et al., 2018), an architecture based on recurrent neural networks, which was followed by transformer neural networks such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Universal Sentence Encoder (USE) (Cer et al., 2018). Deep neural networks of this kind have been successfully applied to many NLP tasks, demonstrating state-of-the-art performance in question answering, natural language inference, sentiment analysis and multiple other areas covered by the GLUE benchmark (Wang et al., 2018).¹

The present study

In the present study we provide the first evaluation of the usefulness of deep neural networks as a computer-based approach for detecting ASD from textual utterances. Deep neural networks were used to classify participants with ASD and with TD based on their narrations produced during a picture book task. We tried several methods not previously examined for this purpose to identify the best performing approach. We compared the obtained results with the results of standardized testing and classifications made by two psychiatrists experienced in ASD diagnosis.

We aimed to investigate how effective deep-learning models for automatic text analysis would prove to be in differentiating people with TD and atypical development, in our case with ASD. The research was exploratory in nature and served as a basis for future studies and translation of advances in the field of artificial intelligence (AI) and NLP into clinical practice.

METHOD

Participants

We recruited 50 participants: 25 with idiopathic ASD (ASD group) and 25 controls with TD (TD group) matched for age, sex and ethnicity, and nearly matched for intelligence quotients (Table 1). The inclusion criteria were: (1) age ≥ 7 years, (2) non-verbal IQ ≥ 80 , (3) Polish as a first and primary language, (4) no hearing, sight and mobility impairments, and (5) for ASD group—prior clinical diagnosis of

TABLE 1 Demographics and clinical characteristics of the sample

	Mean	SD	Range	p-value
<i>Chronological age (years)</i>				
ASD	14.55	5.46	7.1–24.6	
TD	14.38	5.83	7.3–25.3	0.92
<i>Non-verbal IQ</i>				
ASD	109.08	13.04	80–140	
TD	114.64	12.50	80–134	0.13
<i>Verbal IQ</i>				
ASD	106.62	19.70	72–131	
TD	113.12	14.33	85–141	0.20
<i>ADOS-2 overall score</i>				
ASD	13.60	5.23		
TD	2.48	3.22		< 0.001
<i>SCQ overall score</i>				
ASD	22.19	6.01		
TD	3.60	5.78		< 0.001

ASD, autism spectrum disorder; TD, typical development; ADOS-2, Autism Diagnostic Observation Schedule, Second Edition; SCQ, Social Communication Questionnaire; SD, standard deviation; IQ, Intelligence Quotient; p-value of an independent samples *t*-test.

ASD determined by a multidisciplinary team, including a psychiatrist based on ICD-10 diagnostic criteria (WHO, 2009). The exclusion criteria for controls included: (1) a personal or family history of ASD, (2) a history of developmental disorders and (3) neurological or psychiatric conditions or suspected genetic syndromes and developmental problems. Prior to this study, participants were not evaluated using SCQ and ADOS-2 measures.

Measures and procedure

The project was approved by the Faculty of Psychology, University of Warsaw Research Ethics Committee. Informed consent was signed prior to participation in the study by: (1) the parents of participating children under 16 years of age, (2) the parents of participating children aged 16 and older and (3) the participants themselves. All procedures were in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Participants were contacted through university clinic, diagnostic or therapeutic centres, or schools. They were assessed as part of a research evaluation. The ASD symptoms were assessed by a direct observation of a participant using Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) and parent-report on the Social Communication Questionnaire (SCQ). Research-reliable clinicians

conducted the ADOS-2 assessments with the supervision of the present author, who is an ADOS-2 trainer.

The ADOS-2 assessment and picture book narration

ADOS-2 is a standardized, semi-structured observation schedule that includes a number of play-based and conversational activities designed to obtain information concerning social communication skills and restricted and repetitive behaviours associated with a diagnosis of ASD (Lord et al., 2012b,a). The Polish version of the ADOS-2 demonstrates good psychometric characteristics (Chojnicka & Pisula, 2017).

Language samples were derived from the ADOS-2 *Telling a Story from a Book* task. In order to elicit narratives, examiners used the picture book *Tuesday* by David Wiesner. The book depicts the adventures of frogs which one Tuesday evening start to float on the lily pads and fly to the nearest town. The pictures show unreal and humorous situations, and several mental and emotional states of the characters.

The participants' narratives during the ADOS-2 *Telling a Story from a Book* task were video recorded. Two experienced transcribers independently transcribed the narrative produced by a participant, with word-level reliability averaging 91%. The transcribers were blind for diagnosis and followed the Codes for Human Analysis of Transcripts (MacWhinney, 2000). One of the trained transcribers compared prepared transcripts word for word. In case of discrepancies, the transcriber made corrections based on the video recordings.

Table 2 contains basic descriptive statistics of the transcriptions.

On average participants uttered statements related to 13.6 out of 15 pages of the book.

Other measures

The SCQ is a parent questionnaire useful for evaluation of communication skills and social functioning in children over 4.0 years, adolescents and adults who may have ASD (Rutter et al., 2003). The Polish version of the SCQ is characterized by good psychometric properties (Pisula, 2017). For the Overall total score, the intraclass correlation coefficient for test–retest reliability was 0.91, Cronbach's alpha coefficient as an indicator of internal consistency was 0.92. The agreement between the Polish version of SCQ and clinical diagnosis (ASD versus non-spectrum) as well as ADI-R results were excellent: Cohen's kappas = 0.84 and 0.81, respectively (Cohen 1960; Cicchetti & Sparrow, 1981).

TABLE 2 Statistics of transcriptions

	Mean	SD	Range
Number of sentences	22	11	3–60
Number of tokens per narration	323	143	70–720
Duration of the task (h:min:s)	00:04:24	00:01:18	00:02:28–00:07:22

SD, standard deviation.

Cognitive abilities were assessed using the Wechsler Intelligence Scale for Children—Revised for verbal children and adolescents aged 6.0–16.11 (Matczak et al., 2008), and the Wechsler Adult Intelligence Scale for verbal participants older than 16.11 (Brzeziński et al., 2004).

Neural networks in detecting ASD

We compared several methods used to classify subjects with ASD versus TD by using some established methods described in the literature and some novel ones. The approaches we tested differ not only in terms of multiple ways of representing utterances such as ELMo or USE text encoders, but also in the way of handling missing data or classification algorithms.

To assess the model's ability to predict unknown data, we applied a leave-one-out cross-validation. This is a special case of cross-validation where the number of folds equals the number of occurrences (utterances of one person, in our case) in the data set. The training algorithm is applied once for each instance, using all other instances as the training set. This type of cross-validation maximizes the size of the training set (which is an advantage in the case of a relatively small amount of data) at the expense of more repetitions of the training algorithm.

Embeddings from Language Models (ELMo)

The first method we applied to detect ASD from textual utterances is Embeddings from Language Models (Peters et al., 2018). It is a pre-trained, multilayer, bidirectional neural network language model. For each text input, ELMo uses a convolutional neural network layer (CNN) to obtain word-level vectors. It then uses two layers of bidirectional long short-term memory (LSTM) recurrent neural networks (Hochreiter & Schmidhuber, 1997).

Both CNN and LSTM layers in ELMo were pre-trained on large corpora in the language modelling task: predicting the next word in a document using provided context. The language model is first fed a large amount of unannotated data to learn the usage words and acquire linguistic knowledge. Language models pre-trained in such a way

are effective and universal means for encoding meaning of natural language texts. They may be effectively applied to solve target tasks on smaller datasets. In our case, to detect ASD.

The results in our paper used the Polish language ELMo version (Che et al., 2018) provided by ELMoForManyLangs. We used the default settings. Pre-training of ELMoForManyLangs was performed on a 20 million-word data set sampled from Wikipedia and Common Crawl. For each input utterance, the representation obtained from ELMo was a vector of 1024 elements. The vectors were then used as an input to supervised classification algorithms. A simplified schema of encoding and classifying utterances using ELMo can be found on Figure 1.

Universal Sentence Encoder (USE)

The second method to obtain text representations was USE developed by researchers at Google (Yang et al., 2019). This deep neural network text encoder supports 16 languages, among them Polish. The input to the network is variable length text in any of the 16 supported languages and the output is a 512-dimensional vector. The multi-task training setup of USE was based on the technique described by Chidambaram et al. (2018).

We used the *large* variant (version 3) of the USE, which is based on transformer neural network architecture (Vaswani et al., 2017). The context-aware word representations are converted to a fixed-length sentence encoding vector (a vector representing the input text) by computing the element-wise sum of the representations at each word position (Cer et al., 2018). A simplified schema of encoding and classifying utterances using a transformer neural network such as USE is illustrated in Figure 2.

Data representation

We tested several variants of generating representations from transcribed utterances. Option names are marked in **bold** and subsequently used in Table 3 to describe the experimental settings for obtained results.

The first option was to use a **single** vector computed from all utterances of a participant from all pages,

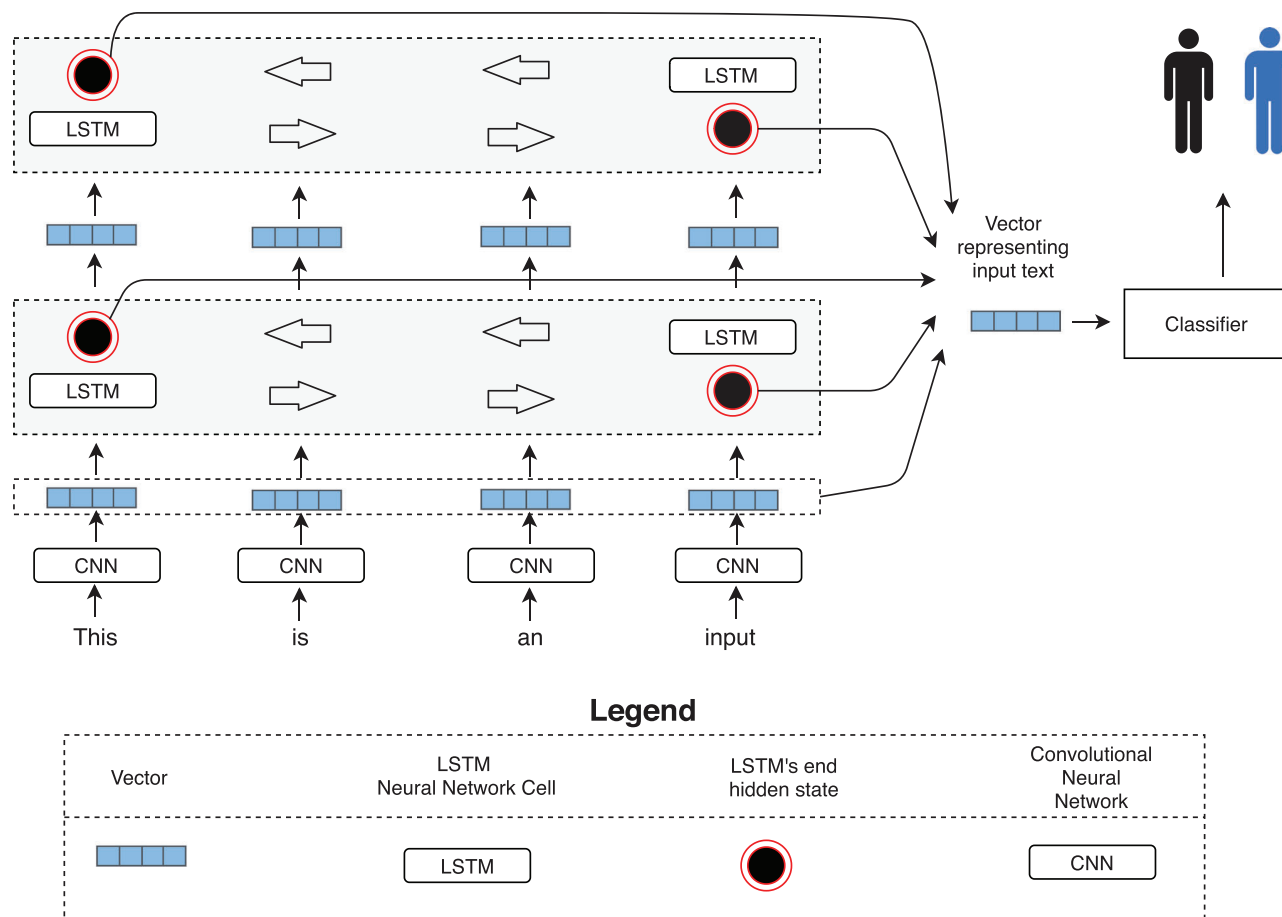


FIGURE 1 Schema of ELMo applied to detecting ASD [Colour figure can be viewed at wileyonlinelibrary.com]

concatenated into one long text using a white character (space). In the second option, embedding vectors were computed for all utterances associated with each **page** in the book. Page-level vectors were then concatenated into one large vector for classification. This scenario needs to address the missing data problem: in about 15% of cases, a participant did not say anything for a specific page in the book while narrating the story and thus an utterance embedding vector could not be created.

We tested two ways of dealing with the missing data problem. The first is to insert vectors of zeros (marked as **0s** in Figure 3 and Table 3) to represent pages with no data. This solution is suitable for machine-learning methods such as XGBoost (Chen & Guestrin, 2016) because it ensures that large vectors obtained from individual page vectors by concatenation are of the same length. In this approach, vector length was 7680 (512×15 pages) in the case of USE and 15,360 (1024×15 pages) in the case of ELMo.

The second solution for the missing data issue is the mechanism called **masking**, specific to neural networks

and originally designed for processing sequence data where individual samples have different lengths. Since the input data for any deep-learning model must be of fixed size, samples that are shorter than the longest item are padded with some placeholder value. Masking is the mechanism that informs the model that some part of the data should be ignored. Because this solution is specific to neural networks, the classification algorithm in this case is based on the Dense layer.²

Data representation methods are illustrated in Figure 3. In this simplified example of four pages (assuming also four as the total number of pages, and thus four vectors as inputs to the classification models), utterances related to the second page did not occur. In the case of **masking**, a special slot labelled MASK is inserted at the end to satisfy the required four-page size. In the **pages** representation with **0s**, the second page has been represented as a vector filled with zeros. Embeddings were then concatenated. In the last case of the **single** representation, the text on three sides was merged into one, and then an embedding vector was computed.

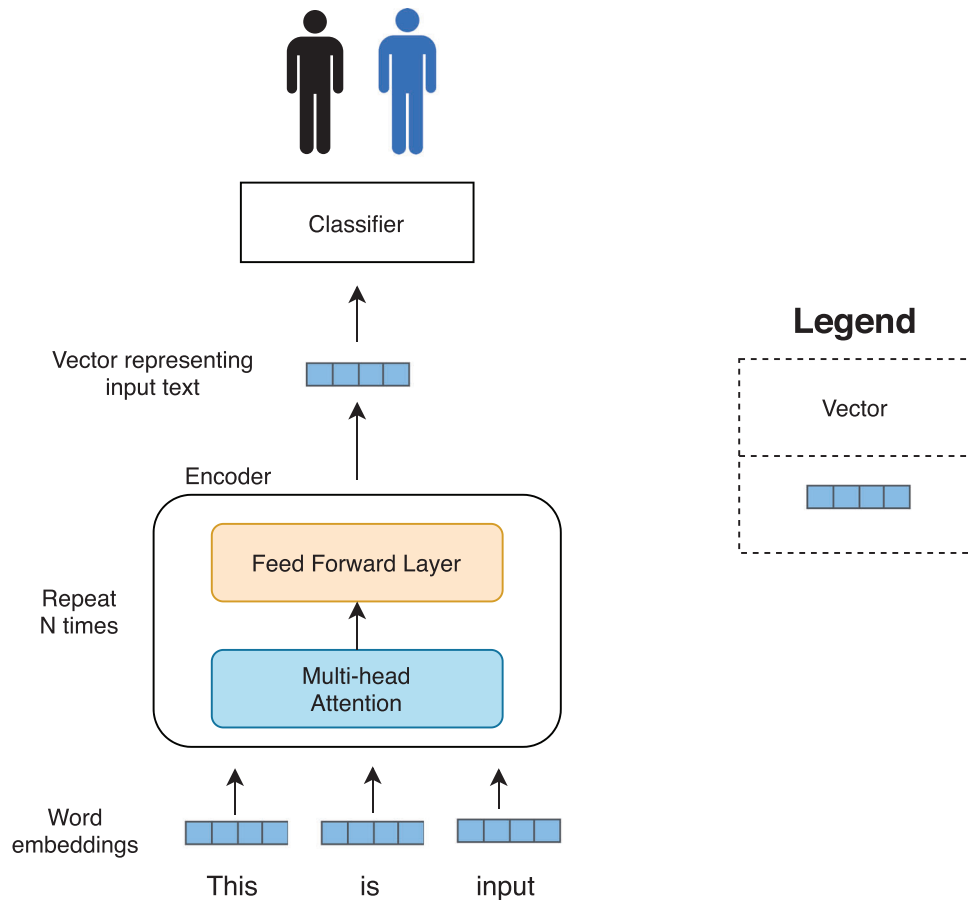


FIGURE 2 Schema of USE applied to detecting ASD [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

Classification algorithms

Classification is an essential step of the analysis that involves predicting which class an item (in our case, a participant) belongs to. In our study the classification was binary: ASD versus TD. For each combination of data representation and text encoder, we tested multiple classification algorithms:

- XGBoost (Chen & Guestrin, 2016).³
- Support vector machines (SVM) with radial (rbf) and linear kernel implementations from the Scikit-Learn package (Pedregosa et al., 2011).⁴
- Single layer Dense neural network with 'relu' activation. We tested the following parameters (ranges reported in parentheses): number of epochs (50, 80, 100), dropout (0.2, 0.3, 0.4) and number of units (128, 256, 512). We used the Tensorflow 2 library (Abadi et al., 2015).

For each input combination we only report the results of the best performing classifier. In case of the Dense clas-

sifier, the only feature space we tested is **masking**, as described above. This relatively simple classifier is the only one with support for **masking**, as no such technique is available for XGBoost or SVM.

Human raters

We also asked two psychiatrists (raters 1 and 2) (Table 3) experienced in the diagnosis of ASD to classify participants based on their transcribed narratives. The raters were asked to read utterances produced by a participant (without the examiner's statements) and assign a participant to one of the two groups, ASD or TD. The raters made the decision based solely on their judgment, without any additional guidance. The scientific goal of this analysis is to examine how a human would interpret the same type of data in comparison with ANNs. We would also like to point out that in a real clinical assessment an experienced clinician does not base her/his decision on such fragmentary data.

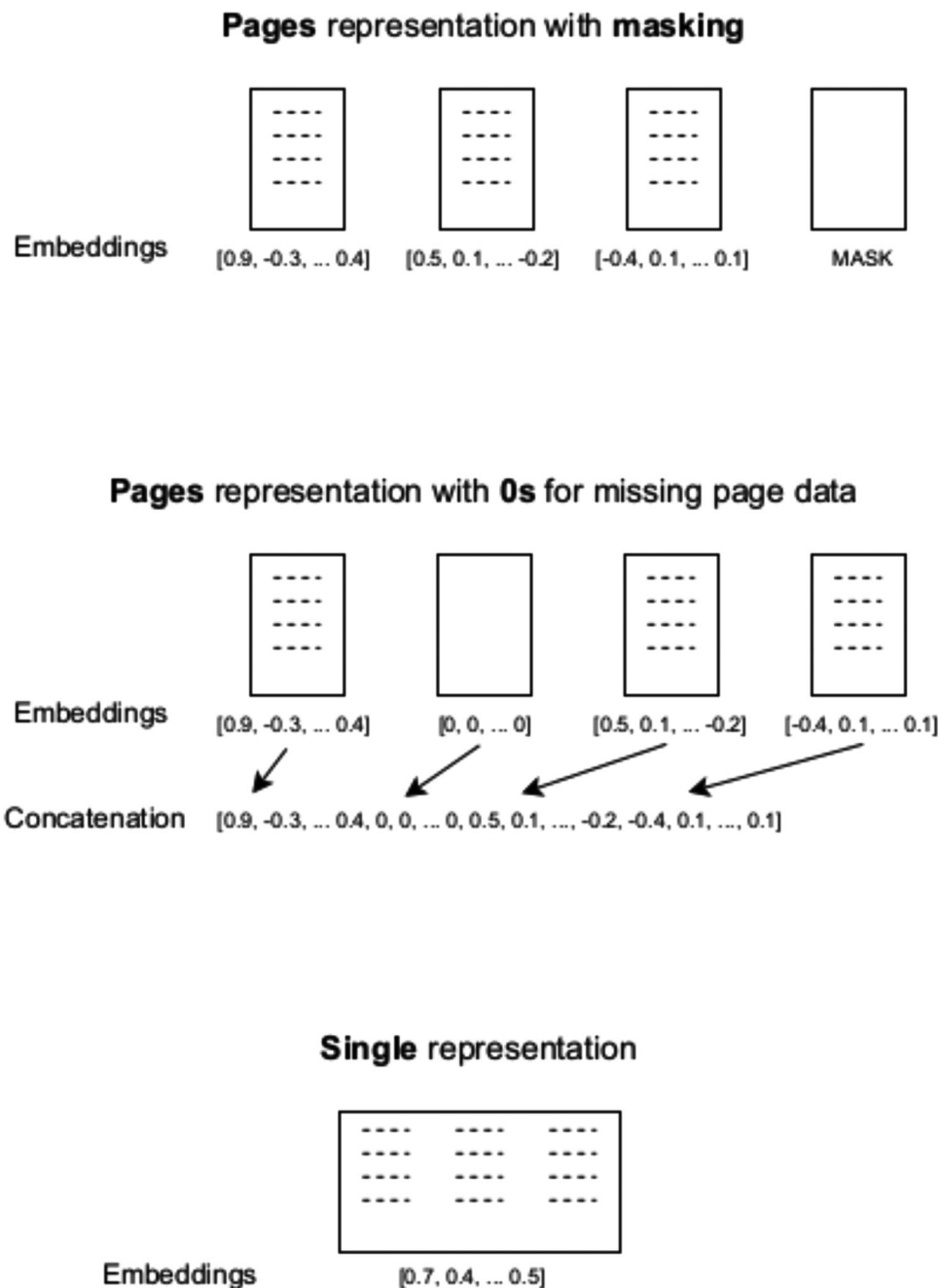


FIGURE 3 Methods of data representation

RESULTS

Table 3 contains the values of sensitivity, specificity, positive (PPV) and negative (NPV) predictive values. In the case of computer-based models (lower part of the table) values were computed in a leave-one-out cross-validation. In the case of raters, they assigned participants to one of the two groups based solely on transcribed utterances, with no additional clinical information.

As shown in Table 3, the two standardized instruments had the highest values of sensitivity, specificity, PPV and NPV. However, the computer-based models performed better than the raters. Both USE and ELMo with the page approach, vectors of zeros for the missing data and XGBoost (USE)/SVM (ELMo) yielded the highest results among computer-based models. For the single vector approach, ELMo + SVM yielded higher results than USE.

TABLE 3 Instruments and methods used for the ASD assessment

	Sensitivity	Specificity	PPV	NPV
ADOS-2	1.00	0.92	0.93	1.00
SCQ	0.92	0.96	0.96	0.92
Rater 1	0.56	0.60	0.48	0.48
Rater 2	0.60	0.64	0.65	0.61
ELMo, single, SVM	0.76	0.60	0.65	0.71
ELMo pages, masking, Dense	0.60	0.56	0.58	0.58
ELMo, pages, 0s, SVM	0.72	0.68	0.69	0.71
USE, single, SVM	0.56	0.60	0.58	0.58
USE, pages, masking, Dense	0.68	0.68	0.68	0.68
USE, pages, 0s, XGBoost	0.72	0.68	0.69	0.71

ADOS-2, Autism Diagnostic Observation Schedule, Second Edition; Dense densely connected neural network layer as a classifier; ELMo, Embeddings from Language Models (Peters et al., 2018); masking = samples shorter due to missing data were padded with a mask; NPV, negative predictive value; pages = page-level embedding vectors (computed for all utterances associated with each page in the book); PPV, positive predictive value; SCQ, Social Communication Questionnaire; single = one embedding vector computed from all utterances; SVM, support vector machines classifier; USE, Universal Sentence Encoder (Yang et al., 2019); XGBoost, gradient boosting classifier (Chen & Guestrin, 2016); 0s = vectors of zeros used to represent pages with no data.

DISCUSSION

This study investigated the ability of several types of deep neural network-based text representation models to detect ASD. Given the well-documented narrative difficulties encountered by individuals with ASD (Baixauli et al., 2016), this study aimed to explore how well different neural networks distinguish participants with ASD and controls with TD based on their utterances produced during a picture book task. We present a novel, exploratory approach not previously described in the literature.

Generally, the best overall results were achieved by the more fine-grained **pages** representation (separate vector generated for utterances related to each page in the book) and zeros (**0s**) to represent missing data (when a participant did not say anything for a specific page in the book while narrating the story). Interestingly, identical top scores were achieved by both ELMo and USE text encoders, which indicates that the text encoding method might be of less importance.

The success of **pages** representation may be attributed to multiple related facts. This representation contains much more information than **single** vectors, since **pages** contains detailed representation of utterances related to specific pages in a book. The increase in the amount of information is reflected in the input of classification models: they are fed with vectors of size 15,360 (ELMo) and 7680 (USE), instead of just 1024 and 512 as in the case of **single** vectors. Apparently, the aggregation that occurs in the **single** vector approach for the whole narration causes some key information to be lost. It is likely that information contained in certain pages may be actually more crucial than in the others and needs to be represented in more detail. Even missing out certain pages may provide clues

to identify ASD participants. Reflecting such phenomena is indeed the best in the case of combined **pages** and **0s** utterance representation, which is an important contribution of our approach. Nonetheless, the analysis that takes into account the division of narration into subsets referring to individual pages has some drawbacks. In our study it was done manually, although one may envision automation of this step. Nevertheless, the division of the whole narrative into subsets associated with individual pages is sometimes ambiguous, especially when a participant skips one of the pages during storytelling. In our study, we used video recordings that allowed us to identify the skipped pages by a participant. This may be more challenging for the audio recordings and may require arbitrary decisions whether the utterance should be assigned to one page or the next. Therefore, in the future it might also be worth developing the approach with a single vector computed from the whole narration produced by a participant. In that case, ELMo encoder appears to be the most promising, with comparable sensitivity and specificity values as page-based approaches. Automated models outperformed experienced psychiatrists on the task of assigning participants to one of the two groups, ASD or TD, using only the utterances produced by a participant. This does not indicate that neural networks can replace clinical judgment. An experienced clinician does not learn, in contrast to the ANNs, to diagnose patients based on their isolated statements. On the contrary, multidisciplinary and multimodal diagnostic assessment should include information from many sources, including observations and assessments on the functioning of a person in various environments and contexts (Goldstein & Ozonoff, 2018). The results appear to confirm that automated ASD classification based on narrative transcriptions is feasible and that its accuracy is higher

than that obtained by a human with the same task. We argue that computational approaches are worth studying and along with other tools may in the future provide support for the clinician analogous to the standardized diagnostic instruments used today.

Conversely, both ADOS-2 and SCQ tools demonstrated diagnostic effectiveness way above the computer models. Both instruments measure not solely language-related characteristics, but communication skills, social functioning and patterns of behaviours and interests. ADOS-2 sensitivity and specificity values obtained in our study were higher than those described in the validation study: ADOS-2 sensitivity values were 90% and 92% and specificity values were 81% and 74% for modules 3 and 4, respectively (Chojnicka & Pisula, 2017). Higher values of sensitivity and specificity in the current study may reflect the fact that all examiners were trained in the use of the ADOS-2 for research and clinical purposes and supervised by an independent ADOS-2 trainer. Findings by Zander et al. (2016) endorse the objectivity of the ADOS in naturalistic clinical settings, yet the sensitivity and specificity values for examiners without research reliability training may be lower. Furthermore, participants in our sample were carefully selected with clear signs of ASD among individuals in the ASD group and lack of developmental problems among individuals in the TD group. In the validation sample, the control group consisted not only of healthy participants, but also individuals with non-spectrum disorders (Chojnicka & Pisula, 2017; Lord et al., 2012b). This might also explain higher values of sensitivity and specificity of SCQ in our study than those reported in the literature (Barnard-Brak et al., 2015; Wiggins et al., 2007). It is worth considering, however, that the assessment using standardized methods, in particular an observation protocol like ADOS-2, is time and labour-consuming. By contrast, possible future implementations of automated tools might require less human labour and expense.

The results obtained in our experiments on detecting ASD from transcribed clinical interviews are comparable with other published attempts at automated ASD detection. Examples of detection from texts include Cho et al. (2019), who use acoustic and text features drawn from short, unstructured conversations. Data were converted to 624 features (352 acoustic + 272 text) to apply a Gradient Boosting classifier. This resulted in an accuracy of 0.75, which is on a par with our results despite the additional acoustic information. Automated ASD detection remains challenging even when using a very different type of data: a convolutional neural network on fMRI brain scans (Sherkatghanad et al., 2020) reaches an accuracy of 0.70.

Limitations and future directions

There are several limitations of our study as well as promising directions for future research. One limitation was a relatively small sample size, with a wide age range of participants and not a perfect match for intelligence quotients. As a consequence, we were not able to divide our sample into subgroups based on age, sex or IQ level. Future works should address this restraint and replicate our findings with larger samples. Particular attention should be paid to the possible sex differences and impact of the IQ on results. It is important to mention that our study included only fluently speaking participants with ASD with no intellectual disability and might not apply to everyone on the autism spectrum. This scenario is more problematic for text-only predictions. From this perspective, the results we have obtained can be considered valuable, as perhaps they show the upper limit of diagnostic possibilities from the closed text data only.

We have tested narratives produced during a picture book task. Previous research suggests that narrative context matters. Participants with ASD exhibit lower performance in less structured tasks (i.e., narratives of personal experiences, conversation) than more structured picture book tasks (Lee et al., 2019; Losh & Capps, 2003). Thus, future research should consider different types of narrative stimuli that may affect the effectiveness of automated methods.

Furthermore, as we compared individuals with ASD and controls with TD, future studies should include individuals with non-spectrum disorders and other psychiatric conditions.

The solution described in our paper requires high-quality speech-to-text transcription. In the case of spontaneous conversations, this is often difficult to achieve using software only. Although automatic speech recognition continues to improve, manual adjustment of transcriptions would still be necessary on data such as we analyse here.

Neural network encoders followed by classifiers can take into account a wide variety of utterance characteristics but does not help to identify which particular aspects of language are crucial. For example, Chojnicka and Wawer (2020) linked ASD to such aspects of language as sentiment and abstraction. However, it is very likely that this list is not exhaustive. The identification of other language characteristics may be possible using the architecture proposed in our paper by applying model explanation techniques to compute the contribution of each word in an utterance to the classifier's decision (Ribeiro et al., 2016). Giving meaning to collections of important words may lead to discovering

new dimensions of language relevant to ASD. We plan to pursue this in the future work.

Conclusions and clinical implications

The current findings elicit some clinical considerations about the speech and language assessment and intervention planning in ASD. Clinical language measures and tests cannot thoroughly capture the communication struggles that people with ASD face (Barokova & Tager-Flusberg, 2020). This study has demonstrated that computational methods detect language differences between narratives produced by individuals with ASD and with TD that are less effectively recognized by human raters. In the future, we plan to explore the specific language characteristics that are most relevant for deep-learning models and classification decisions (explainable AI approach). Studying pragmatic language ability in ASD and developing automated, objective measurements may assist clinicians during the screening and diagnosis, especially for participants with less severe autism spectrum-related symptoms, more advanced language, and no intellectual disability. It can also help design meaningful intervention strategies and adjust educational and therapeutic programs (Mertz, 2021).

The current study demonstrated the effectiveness of automated ASD detection using two state-of-the-art deep-learning models, ELMo and USE. We investigated predictions from textual utterances, insufficiently researched to date in the field of neurodevelopmental disorders. Both ELMo and USE text encoders provided promising values of specificity, sensitivity, PPV and NPV. The single vector approach, rather than the pages approach, seems to have the greatest possible clinical utility. While we have achieved a slightly higher specificity of PPV and PPV using the pages approach, subdividing the narrative into subsets can be problematic and sometimes ambiguous. In contrast, the use of an encoder with a single vector approach enables full automation and objectification of the measurement.

We focused on text as input, but it is conceivable to combine it with speech functions and extra-linguistic variables (Asgari et al., 2021; Fusaroli et al., 2017). Our results pave the way to further research applying model explanation techniques and including other study groups and narrative contexts. In the future, the effectiveness of the approach presented in our article may be significantly increased if recorded speech transcriptions can be generated using the currently intensively developed automatic speech-to-text services.

While the results are preliminary and require further research, they do offer promise for the further develop-

ment of next-generation solutions that can support clinicians and augment the decision-making process associated with ASD diagnosis, screening, and intervention planning. Future work should aim to refine the use of deep-learning models for text analysis that have a high potential for translation into clinical practice.

ACKNOWLEDGEMENT

Both the authors contributed equally to this paper.

CONFLICT OF INTEREST

The authors declare that there are no competing interests.


DATA AVAILABILITY HEAD

The data that support the findings of this study are available from authors Aleksander Wawer and Izabela Chojnicka upon request. The data are not publicly available due to them containing information that could compromise the privacy of the research participants.

FUNDING

The study was supported by a project of the National Science Center of Poland (grant number #2020/39/D/HS6/00809) and by the Faculty of Psychology, University of Warsaw, from the funds awarded by the Ministry of Science and Higher Education in the form of a subsidy for the maintenance and development of research potential in 2022 (grant number 501-D125-01-1250000 zlec*. 5011000226).

ORCID

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

Izabela Chojnicka  <https://orcid.org/0000-0001-8723-6873>

NOTES

¹The most recent list of top performing solutions is accessible at <https://gluebenchmark.com/leaderboard>.

²We also experimented with several variants of recurrent neural networks, but the results were far from satisfactory and substantially worse than those obtained using the Dense layer.

³We used the XGBoost python package and API (<https://xgboost.readthedocs.io/en/stable/>) with the default settings.

⁴We used the NuSVC class and implementation.

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kud-Lur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu,

- Y. & Zheng, X., (2015) TensorFlow: large-scale machine learning on heterogeneous systems. <https://www.tensorflow.org/>.
- Asgari, M., Chen, L. & Fombonne, E., (2021) Quantifying voice characteristics for detecting autism. *Frontiers in Psychology*, 12, 2805.
- Baixauli, I., Colomer, C., Roselló, B. & Miranda, A., (2016) Narratives of children with high-functioning autism spectrum disorder: a meta-analysis. *Research in Developmental Disabilities*, 59, 234–254.
- Bal, V.H., Kim, S.H., Cheong, D. & Lord, C., (2015) Daily living skills in individuals with autism spectrum disorder from 2 to 21 years of age. *Autism*, 19(7), 774–784.
- Barnard-Brak, L., Brewer, A., Chesnut, S., Richman, D. & Schaeffer, A.M., (2015) The sensitivity and specificity of the social communication questionnaire for autism spectrum with respect to age. *Autism Research*, 9(8), 838–845.
- Barokova, M. and Tager-Flusberg, H., (2020) Commentary: measuring language change through natural language samples. *Journal of Autism and Developmental Disorders*, 50(7), 2287–2306.
- Brzeziński, J., Gaul, M., Hornowska, E., Jaworowska, A., Machowski, A. & Zakrzewska, M., (2004) WAIS-R (PL)—Wechsler Adult Intelligence Scale — a revised form. Renormalization 2004 (Psychological Test Laboratory of the Polish Psychological Association).
- Capps, L., Losh, M. & Thurber, C., (2000) “The frog ate the bug and made his mouth sad”: narrative competence in children with autism. *Journal of Abnormal Child Psychology*, 28(2), 193–204.
- Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B. & Kurzweil, R., (2018) Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Association for Computational Linguistics, Brussels, Belgium), 169–174. URL: <https://www.aclweb.org/anthology/D18-2029>.
- Che, W., Liu, Y., Wang, Y., Zheng, B. & Liu, T., (2018) Towards better UD parsing: deep contextualized word embeddings, ensemble, and treebank concatenation. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Association for Computational Linguistics, Brussels, Belgium), 55–64. <http://www.aclweb.org/anthology/K18-2005>.
- Chen, T. and Guestrin, C., (2016) XGBoost: a scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (ACM, New York, NY, USA), 785–794.
- Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y., Strope, B. & Kurzweil, R., (2018) *Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model*. CoRR, abs/1810.12836. <http://arxiv.org/abs/1810.12836>.
- Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R.T. & Parish-Morris, J., (2019) Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. In: G. Kubin and Z. Kacic, editors, Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019 (ISCA), 2513–2517.
- Chojnicka, I. and Pisula, E., (2017) Adaptation and Validation of the ADOS-2, Polish Version. *Frontiers in Psychology*, 8.
- Chojnicka, I. and Wawer, A., (2020) Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. *Plos One*, 15(3).
- Cicchetti, D.V. and Sparrow, S.A., (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86(2), 127–137.
- Cohen, J., (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Devlin, J., Chang, M., Lee, K. & Toutanova, K., (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. CoRR, abs/1810.04805. <http://arxiv.org/abs/1810.04805>.
- Diehl, J.J., Bennetto, L. & Young, E.C., (2006) Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 34(1), 83–98.
- Engberg-Pedersen, E. and Christensen, R.V., (2016) Mental states and activities in Danish narratives: children with autism and children with language impairment. *Journal of Child Language*, 44(5), 1192–1217.
- Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D.M. & Gaigg, S.B., (2017) “Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis”. *Autism Research*, 10(3), 384–407.
- Goldstein, S. and Ozonoff, S., (2018) *Assessment of Autism Spectrum Disorder*, 2nd ed. (The Guilford Press).
- Gustafsson, L. and Papliński, A.P., (2004) Self-organization of an artificial neural network subjected to attention shift impairments and familiarity preference, characteristics studied in Autism. *Journal of Autism and Developmental Disorders*, 34(2), 189–198.
- Hochreiter, S. and Schmidhuber, J., (1997) Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kim, S.H. and Lord, C., (2012) Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Child Psychology and Psychiatry*, 53(2), 143–151.
- Lai, M.C., Lombardo, M.V. & Baron-Cohen, S., (2014) *Autism. The Lancet*, 383(9920), 896–910.
- Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y. & Cheng, G., (2020) A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks*, 122, 338–363.
- Lee, M., Martin, G.E., Hogan, A., Hano, D., Gordon, P.C. & Losh, M., (2018) What’s the story? A computational analysis of narrative competence in autism. *Autism*, 22(3), 335–344.
- Lee, M., Nayar, K., Maltman, N., Hamburger, D., Martin, G.E., Gordon, P.C. & Losh, M., (2019) Understanding social communication differences in autism spectrum disorder and first degree relatives: a study of looking and speaking. *Journal of Autism and Developmental Disorders*.
- Lord, C., Luyster, R.J., Gotham, K. & Guthrie, W., (2012a) Autism diagnostic observation schedule, *Second Edition (ADOS2) Manual (Part II): Toddler Module (WPS)*.
- Lord, C., Rutter, M., Dilavore, P.C., Risi, S., Gotham, K. & Bishop, S.L., (2012b) Autism diagnostic observation schedule, Second Edition (ADOS-2) *Manual (Part I): Modules 1–4 (WPS)*.
- Losh, M. and Capps, L., (2003) Narrative ability in high-functioning children with Autism or Asperger’s syndrome. *Journal of Autism and Developmental Disorders*, 33(3), 239–251.
- Losh, M., Childress, D., Lam, K. & Piven, J., (2008) Defining key features of the broad autism phenotype: a comparison across parents of multiple and single-incidence autism families. *American*

- Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147B(4), 424–433.
- Losh, M. and Gordon, P.C., (2014) Quantifying narrative ability in autism spectrum disorder: a computational linguistic analysis of narrative coherence. *Journal of Autism and Developmental Disorders*, 44(12), 3016–3025.
- Losh, M., Klusek, J., Martin, G.E., Sideris, J., Parlier, M. & Piven, J., (2012) Defining genetically meaningful language and personality traits in relatives of individuals with fragile X syndrome and relatives of individuals with autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159B(6), 660–668.
- Macwhinney, B., (2000) *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.) (Erlbaum).
- Maenner, M.J., Shaw, K.A., Baio, J., Washington, A., Patrick, M., Dirienzo, M., Christensen, D.L., Wiggins, L.D., Pettygrove, S., Andrews, J.G. & ET AL., (2020) Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, (2016) *MMWR. Surveillance Summaries*, 69(4), 1–12.
- Matczak, A., Piotrowska, A. & Ciarkowska, W., (2008) WISC-R–The Wechsler Intelligence Scale for Children–A Revised Edition (Psychological Test Laboratory of the Polish Psychological Association).
- Mertz, L., (2021) AI, virtual reality, and robots advancing autism diagnosis and therapy. *IEEE Pulse*, 12(5), 6–10.
- Mäkinen, L., Loukusa, S., Leinonen, E., Moilanen, I., Ebeling, H. & Kunnari, S., (2014) Characteristics of narrative language in autism spectrum disorder: evidence from the Finnish. *Research in Autism Spectrum Disorders*, 8(8), 987–996.
- Mäntylä, M.V., Graziotin, D. & Kuuttila, M., (2018) The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27, 16–32.
- Park, J., Ichinose, K., Kawai, Y., Suzuki, J., Asada, M. & Mori, H., (2019) Macroscopic cluster organizations change the complexity of neural activity. *Entropy*, 21(2), 214.
- Parsons, L., Cordier, R., Munro, N., Joosten, A. & Speyer, R., (2017) A systematic review of pragmatic language interventions for children with autism spectrum disorder. *Plos One*, 12(4).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., (2011) Scikit-Learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L., (2018) Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (Association for Computational Linguistics), 2227–2237. <http://aclweb.org/anthology/N18-1202>.
- Philippsen, A. and Nagai, Y., (2018) Understanding the cognitive mechanisms underlying autistic behavior: a recurrent neural network study. In: 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob) (IEEE). <https://doi.org/10.1109/DEVLRN.2018.8761038>.
- Pisula, E., (2017) N.D. Unpublished manuscript.
- Ribeiro, M.T., Singh, S. & Guestrin, C., (2016) "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016. 1135–1144.
- Rutter, M., Bailey, A. & Lord, C., (2003) *The Social Communication Questionnaire Manual* (WPS).
- Sah, W.H. and Torng, P.C., (2015) Narrative coherence of Mandarin speaking children with high-functioning autism spectrum disorder: an investigation into causal relations. *First Language*, 35(3), 189–212.
- Sampaio, F., Feldman, I., Lavelle, T.A. & Skokauskas, N., (2021) The cost-effectiveness of treatments for attention deficit-hyperactivity disorder and autism spectrum disorder in children and adolescents: a systematic review. *European Child & Adolescent Psychiatry*.
- Schaaf, C.P., Betancur, C., Yuen, R.K.C., Parr, J.R., Skuse, D.H., Gallagher, L., Bernier, R.A., Buchanan, J.A., Buxbaum, J.D., Chen, C.A. & et al. (2020) A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nature Reviews Genetics*, 21(6), 367–376.
- Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U.R., Khosrowabadi, R. & Salari, V., (2020) Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience*, 13, 1325.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U. & Polosukhin, I., (2017) Attention is all you need. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems* 30 (Curran Associates, Inc.), 5998–6008. URL attention-is-all-you-need.pdf.
- Wallace, S., Fein, D., Rosanoff, M., Dawson, G., Hossain, S., Brennan, L., Como, A. & Shih, A., (2012) A global public health strategy for autism spectrum disorders. *Autism Research*, 5(3), 211–217.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S., (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP (Association for Computational Linguistics, Brussels, Belgium), 353–355. <https://aclanthology.org/W18-5446>.
- Wawer, A. and Sarzyńska, J., (2018) The Linguistic Category Model in Polish (LCM-PL). In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (European Language Resources Association (ELRA), Miyazaki, Japan). <https://www.aclweb.org/anthology/L18-1696>.
- Wiggins, L.D., Bakeman, R., Adamson, L.B. & Robins, D.L., (2007) The utility of the social communication questionnaire in screening for autism in children referred for early intervention. *Focus on Autism and Other Developmental Disabilities*, 22(1), 33–38.
- World Health Organization, (2009) *International Statistical Classification of Diseases and Related Health Problems: ICD-10*.
- World Health Organization, (2021) *International Statistical Classification of Diseases and Related Health Problems*. <https://icd.who.int/browse11/l-m/enofsubordinateddocument>. Accessed: 23 June 2021.
- Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., Strophe, B. & Kurzweil,



- R., (2019) Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:1907.04307 [cs.CL] <https://arxiv.org/abs/1907.04307>
- Zander, E., Willfors, C., Berggren, S., Choque-Olsson, N., Coco, C., Elmund, A., Moretti, Å.H., Holm, A., Jifält, I., Kosieradzki, et al., (2016) The objectivity of the Autism Diagnostic Observation Schedule (ADOS) in naturalistic clinical settings. *European Child & Adolescent Psychiatry*, 25(7), 769–780.

How to cite this article: Wawer, A. & Chojnicka, I. (2022) Detecting autism from picture book narratives using deep neural utterance embeddings. *International Journal of Language & Communication Disorders*, 57, 948–962. <https://doi.org/10.1111/1460-6984.12731>