­國 立 清 華 大 學

碩 士 論 文

使用自然語言處理從兒童圖畫書的敘述中進行自閉症傾向之分析

Using natural language processing to analyze autistic tendencies from children's picture book narratives

系別：資訊工程學系

學號：109062469

研 究 生：孫睿菡 (Ruihan Sun)

指導教授：陳良弼 (Arbee L.P. Chen) 教授

中 華 民 國 一百一十二 年 七 月

# 摘要

自閉症兒童患者數量不斷增加，但他們未得到社會廣泛關注和理解。根據世界衛生組織估計，全球每100名兒童中大約有1名患有自閉症。這些兒童的自閉症特徵可能在早期就被發現，但通常要到較晚的階段才能得到診斷。此外，不同患者之間的自閉症特徵個體差異極大，對現代醫學構成挑戰。自閉症的診斷過程冗長而複雜，大多數國家和地區都需要專業人員使用各種評估量表和面談來進行診斷。過去的研究指出，自閉症兒童在社交環境中使用語言的能力存在缺陷，例如在講故事方面的技巧。然而，以往的語言能力分析主要依賴於統計分析數據的方法。隨著計算機語言模型的發展，我們期望利用機器學習和深度神經網絡模型對兒童的敘述進行快速的自閉症傾向分析和語言能力分析。這種方法將提高診斷效率，並提供更全面的語言評估，以更好地了解和支援自閉症兒童的需求。

在本論文中，我們進行了一項實驗，針對7名自閉症兒童和16名典型發展兒童，利用兩本風格不同的圖畫書收集他們對圖畫書的敘述。隨後，我們將這些敘述轉錄成文字形式，並利用三種計算機模型架構進行訓練和分類。由於我們所收集到的數據量有限，我們也探討了在數據擴增的情況下模型的性能表現。同時，我們還對不同模型在兩本圖畫書上的表現進行了比較。值得注意的是，我們在模型訓練過程中引入了語言特徵，並與先前研究進行了比較。結果顯示，我們的模型在準確率、敏感度和特異性方面表現出色，均超過了90%，相比之前的研究提升了20%。這些結果展示了我們模型的卓越性能和潛在應用的價值。

# Abstract

Children with autism are a growing population, yet they have not received widespread attention and understanding from the society. According to the World Health Organization, approximately 1 in 100 children worldwide is diagnosed with autism. The characteristics of autism in these children may be identified in early stages, but diagnosis is often delayed until later stages. Additionally, there are individual variations in autism traits, posing significant challenges to modern medicine. The diagnostic process for autism is lengthy and complex, requiring professionals to use various assessment scales and interviews in most countries and regions. Previous research has indicated that children with autism exhibit language impairments in social settings, such as difficulties in storytelling skills. However, past studies mainly relied on statistical analysis of language abilities. With the advancement of computational language models, we aim to utilize machine learning and deep neural network models for rapid analysis of children's narratives to detect autism tendencies and assess language abilities. This approach can contribute to improving diagnostic efficiency and providing comprehensive language evaluations to better understand and support the needs of children with autism.

In this paper, we conducted an experiment involving 7 children with autism and 16 typically developing children. We collected their narratives on two different picture books with distinct styles. Subsequently, we transcribed these narratives into written text and trained and classified them using three computational model architectures. Due to the limited amount of data collected, we also explored the performance of the models under data augmentation. Furthermore, we compared the performance of different models on the two picture books. It is worth noting that we incorporated language features into the model training process and compared them with previous studies. The results demonstrated that our model achieved excellent performance, with accuracy, sensitivity, and specificity all surpassing 90%. Furthermore, it exhibited a 20% improvement compared to previous research. These findings highlight the superiority and potential application value of our model.

# Acknowledgment

行文至此，我的研究生時光也即將敲響結束的鐘聲，這基本也會是我學生生涯的終點。此時此刻，我想對所有關心、幫助、支持過我的老師、同學、朋友和家人表示我最真摯的感謝。

我的研究生生涯伴隨著COVID-19的反覆，因為疫情我錯過了大學的畢業典禮，因為疫情出入境需要隔離，我第一次近一年半才回家。雖然疫情改變了很多，但不變的是做研究的態度，這是我從我的導師陳良弼教授身上深深體會到的。我很榮幸能夠成為您的學生，您對學術堅持嚴謹的態度，在每週的會議上，小到報告措辭的合理性大到論文能帶給我們的思考都給予了悉心指導。在研究之外，您對我生活上的照顧與關懷，也都使我感到無比的親切，就像您所說的，您就是一個fixer，可以為大家解決各種問題～

感謝新竹市東區竹蓮國民小學和社團法人新竹市自閉症協進會與本研究的合作，如不是他們提供參與者與實驗場地就無法保證實驗的順利進行。

另外，也要感謝在我研究或是生活上給予過我各種幫助的同學、朋友們。感謝翎愷、昀叡，大家一起策劃實驗，一起與竹蓮國小對接，一起去學校、協會進行實驗以及之後對我的研究提出非常有建設性的意見等等，當然我也不會忘了和大家一起吃飯，玩遊戲，出去玩的快樂時光。當然除了實驗室的同學外，還有很多雖然不是同領域但是在我的學習及生活上默默鼓勵我的朋友。首先就是我的閨蜜鄭穎，雖然她從未有機會來過台灣，但我們還是十年如一日的分享著彼此的日常，也是你總在我心灰意冷的時候給予我堅定的肯定。還有我的球友們，你們從不嫌棄我球技菜，總拉著我打球，甚至還帶我參加了畢業杯，讓我躺著成為了第二名。

最後要感謝我的家人，若不是你們堅定的支持，就沒有現在的我。

To my parents and grandparents who gave me my life and strength,

to my friends who accompanied me through the last 24 years,

and to Shu Chang, who completed my soul.

# Table of Contents

[摘要 i](#_Toc142607273)

[Abstract ii](#_Toc142607274)

[Acknowledgment iv](#_Toc142607275)

[Table of Contents vi](#_Toc142607276)

[List of Tables viii](#_Toc142607277)

[List of Figures ix](#_Toc142607278)

[1. Introduction 1](#_Toc142607279)

[1.1. Overview 1](#_Toc142607280)

[1.2. Narrative ability in ASD 4](#_Toc142607281)

[1.3. Goal 5](#_Toc142607282)

[2. Related Work 6](#_Toc142607283)

[2.1. The Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) assessment and picture book task 6](#_Toc142607284)

[2.2. Child Language Data Exchange System 7](#_Toc142607285)

[2.3. Chinese Language Sample Analysis 9](#_Toc142607286)

[2.4. Computational models for analyzing narration 15](#_Toc142607287)

[3. Data Collection 17](#_Toc142607288)

[3.1 Participants 17](#_Toc142607289)

[3.2 Procedure 18](#_Toc142607290)

[3.3 Materials 19](#_Toc142607291)

[3.4 Get recording and transcription 20](#_Toc142607292)

[3.5 Verbal productivity measures 21](#_Toc142607293)

[3.6 Word analysis 23](#_Toc142607294)

[4. Method 26](#_Toc142607295)

[4.1. Data augmentation 26](#_Toc142607300)

[4.2. Neural networks in detecting ASD 28](#_Toc142607301)

[4.2.1. Support Vector Machines (SVM) 28](#_Toc142607302)

[4.2.2. Bidirectional Encoder Representations from Transformers (BERT) 30](#_Toc142607303)

[5. Experiments 33](#_Toc142607304)

[5.1. Data preprocessing 33](#_Toc142607305)

[5.2. Model settings 34](#_Toc142607306)

[5.3. How much augmentation? 34](#_Toc142607307)

[5.4. Comparison of the learning ability of autism characteristics between two models 37](#_Toc142607308)

[5.5. Is hand coding and word analysis helpful to the model? 40](#_Toc142607309)

[5.6. Comparison with previous study 41](#_Toc142607310)

[6. Conclusion 43](#_Toc142607311)

[7. Limitations and future work 45](#_Toc142607312)

[Reference 47](#_Toc142607313)

# List of Tables

[Table 1: Group descriptive characteristics. 18](#_Toc142607228)

[Table 2: Narrative performance on verbal productivity measures 23](#_Toc142607229)

[Table 3: Word analysis 25](#_Toc142607230)

[Table 4: Data augmentation operation 27](#_Toc142607231)

[Table 5: Verbal productivity and words analysis performance 41](#_Toc142607232)

[Table 6: Our methods compare with (Wawer & Chojnicka, 2022) 42](#_Toc142607233)

# List of Figures

[Figure 1: The flow chart of our experiment 17](#_Toc142607238)

[Figure 2: Snapshots of the two selected picture books. Left side is from Tuesday, and Right side is from *子兒，吐吐* 20](#_Toc142607239)

[Figure 3: Visualization of augmented datasets for picture book *Tuesday* 28](#_Toc142607240)

[Figure 4: TF-IDF with SVM architecture 30](#_Toc142607241)

[Figure 5: TinyBERT with hand coding and word analysis architecture 32](#_Toc142607242)

[Figure 6: Performance on SVM architecture with differentamplification factors 36](#_Toc142607243)

[Figure 7: Performance on TinyBERT architecture with differentamplification factors. 37](#_Toc142607244)

[Figure 8: Two models’ performance on Book 1 for training and Book 2 for testing. 38](#_Toc142607245)

[Figure 9: Two models’ performance on Book 2 for training and Book 1 for testing. 39](#_Toc142607246)

# Introduction

## Overview

Autism, also known as Autism Spectrum Disorders (ASD), encompasses a wide range of brain development conditions. Its characteristics include enduring difficulties in social communication and interaction, and repetitive and restricted behavior patterns (American Psychiatric Association, 2022). Although these characteristics can be observed in early childhood, autism is frequently not diagnosed until later stages. Global estimates suggest that about 1 in 100 children are affected by autism (Zeidan, et al., 2022). Nevertheless, it is important to note that reported prevalence varies significantly among different studies. Autistic individuals exhibit diverse abilities and needs that can change over time. While some can live independently in the community, others face profound challenges and require lifelong care and support. In addition, autistic individuals often encounter challenges in obtaining adequate educational and equal employment opportunities.

According to the autism prevalence report of the Autism and Developmental Disabilities Monitoring (ADDM) Network, the prevalence of ASD in 2020 increased to 27.6 per 1,000 8-year-old children across the United States. The estimate means approximately one in every 36 children is diagnosed with ASD. Moreover, the study unveiled a notable gender disparity, with the disorder being 3.8 times more prevalent in boys than girls (Maenner, et al., 2023). In 2023, the latest prevalence of autistic people in Taiwan is 19,078, with 74.6% identified with mild autism. Most children with mild autism are often detected and receive the diagnosis until they enter the school system. Among all the people identified as ASD in Taiwan, children aged 6-12 are the second highest population across different age ranges (衛生福利部, 2023). The phenomenon indicates that many children are diagnosed with ASD during their school age. These autistic children may encounter a variety of learning and social challenges and need individualized educational support. Therefore, helping school teachers detect children with autistic traits may accelerate the process of providing appropriate education resources.

While there are well-established standardized measurements for assessing ASD characteristics, the process of detecting and confirming the ASD diagnosis is still lengthy and time-consuming. One potential reason for the long diagnostic process may be that the definition of ASD is mainly relied on behavior, not genetic variants or physical biomarkers (Schaaf, et al., 2020). The diagnostic procedure of ASD often involves multifaceted evaluation that consists of two main steps: screening and comprehensive diagnostic evaluation. A typical diagnostic process of ASD usually comprises multidisciplinary collaboration and acquiring information from various sources, including but not limited to: (1) interviews with parents or caregivers, (2) educational diaries or behavioral observation from kindergarten or school teachers, (3) standardized behavioral assessments, and physical examinations (Goldstein & Ozonoff, 2018). The whole diagnostic process often takes up to months with various hospital visits.

Additionally, to be qualified for special education, autistic students are required to go through at least three steps, discovery, screening and evaluation, to obtain their educational ASD identity (教育部, 2014). Take evaluation as an example, the step of evaluation for a single child alone requires psychological-assessment teachers and multiple school-based rehabilitation professionals (such as occupational therapists, speech therapists, and psychiatrists) to conduct interviews, observations, and various standardized assessments . The above assessments may take 30 minutes to three hours, which is time-consuming for children, caregivers, and teachers.

To accelerate the ASD evaluation process, we aimed to advance the Natural Language Processing (NLP) models to support detecting autistic children in the school context. As a computational model, NLP has the potential to offer clinician decision-making support akin to psychological testing tools. For example, in suicidal ideation detection: (Tadesse, Lin, Xu, & Yang, 2019) combined Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to do text classify, on Reddit social media dataset, and reached 93.8% accuracy; (Coppersmith, Leary, Crutchley, & Fine, 2018) used pretrained GloVe embeddings (Pennington, Socher, & Manning, 2014) and LSTM to get the 85% sensitivity. In predicting depression via social media: (De Choudhury, Gamon, Counts, & Horvitz, 2013) used NLP to analyze the linguistic and behavioral characteristics of tweets from both groups. They focused on specific linguistic cues, such as emotional expressions, self-disclosure, and social interactions, that may be indicative of depressive states. And their model achieved average accuracy of ~70%, high precision of 74%.

Using computerized approaches, clinicians and teachers may make more accurate decisions and provide speedy diagnoses with a user-friendly interface. However, as diagnosing ASD is a multidimensional process that includes professional knowledge, experience, and clinical reasoning, the proposed automated analysis of NLP may only be used as a supplementary or supporting tool for pediatricians' diagnosis of ASD.

## Narrative ability in ASD

One of the significant characteristics of autistic individuals is the continuing challenges in social communication and interaction across different social contexts (American Psychiatric Association, 2022). While autistic individuals exhibit a broad range of verbal-linguistic abilities, a notable aspect of their language profile is a pervasive deficit in pragmatic language skills. (Parsons, Cordier, Munro, Joosten, & Speyer, 2017). Pragmatic language skills include the ability to (1) use language to deliver different meanings and purposes, (2) modify language according to the conversation counter partners and the social contexts, (3) understand non-literal languages, such as sarcasm and slang, and (4) follow the conversation rules (Parsons, Cordier, Munro, Joosten, & Speyer, 2017). Across different languages, autistic individuals have been reported consistently to have deficits in understanding and socially using languages (Engberg-Pedersen & Christensen, 2016; Sah & Torng, 2015; Mäkinen, et al., 2014).

Autistic children’ problems with pragmatic language use are most frequently seen in their narratives, such as story-telling (Baixauli, Colomer, Roselló, & Miranda, 2016). When examing the narratives produced by autistic children, researchers found that autistic children’s narratives tend to lack coherence and causal connections, and may include irrelevant or inappropriate components (Diehl, Bennetto, & Young, 2006). In addition, autistic children have fewer utterances and less lexical diversity than their typically developing peers when telling stories (Capps, Losh, & Thurber, 2000). The mindblindness theory of autism may be a possible reason for the challenges in producing comprehensive narratives (Baron-Cohen, Leslie, & Frith, 1985). Autistic children may be unable to describe characters' thoughts and feelings because they do not understand the motivations behind the characters' actions in the story. Due to the narrative characteristics of ASD, we may be able to identify children who possess autistic traits by examing children’s narratives.

## Goal

In our study, we evaluated the validity of NLP as a computer-based method for analyzing autism spectrum disorder tendencies from textual discourse. We tested several previously unexplored approaches and selected the ones that performed best. We aimed to (1) distinguish between autistic children (ASD) and their typically developing peers (TD) through an NLP model of automatic text analysis, (2) explore the linguistic features of the two populations, and (3) improve the model capability by adding the external knowledge of the language features.

# Related Work

In this section, we introduced some relevant tools used in our study and previous studies on the narratives produced by autistic individuals.

## The Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) assessment and picture book task

The ADOS-2 is a standardized assessment tool designed to assist in diagnosing ASD for individuals from 12 months to adulthood (Lord, et al., 2012). In terms of reliability, the ADOS-2 inter-rater reliability is high, and test-retest reliability ranges from .68 to .92 (李宜融, 2015). The ADOS-2 consists of semi-structured activities (including plays and interviews) to provide examiners the opportunities to observe behaviors that are relevant to the diagnosis of ASD. Examiners may choose from five different modules based on an individual's age and communication levels. The assessment takes approximately 40 to 60 minutes to complete. Below are the five modules of the assessment. Only one module would be chosen for a single individual.

* Toddler Module – for children aged 12 to 30 months who do not often use phrases to speak.
* Module 1 – for children 31 months and older who do not use phrases frequently.
* Module 2 – for children of any age who use phrases to speak but are not fluent in spoken language.
* Module 3 – for children and teenagers who are fluent in spoken language.
* Module 4 – for older adolescents and adults who are fluent in spoken language.

In this study, we used the story-telling activity in Module 3 to collect children’s narratives. The activity requires the examed child to tell a story based on a picture book without words: *Tuesday* (Wiesner, 1991).

The story is about the adventures of a group of frogs who float on water lily petals and visit a nearby town. These pictures depict unreal and humorous scenarios, as well as various psychological and emotional states of the characters. Participants were asked to look at pictures and tell stories. The instructions given are as follows: "*Look at this book. It presents a story about frogs. Can you tell me the story as we progress?*" During the data collection, the experimenter tried to avoid getting involved, showing non-verbal expressions of hints or encouragement, or providing guidance. The experimenter's goals were to videotape child’s narratives and transcribe the story using the Codes for Human Analysis of Transcripts (CHAT) (MacWhinney & Snow, 1990).

This task evaluates the children's capacity to comprehend and discuss a sequential story presented in a picture book and utilize visual cues provided to construct a narrative. Additionally, it evaluates the children's aptitude in recounting a story in a sequential manner.

## Child Language Data Exchange System

Established in 1984 by Brian MacWhinney and Catherine Snow, the Child Language Data Exchange System (CHILDES) is the largest computerized database for child languages (MacWhinney & Snow, 1990). The CHILDES has the following three features: (1) A database currently consists of 230 corpora with 30 languages and includes transcripts of spontaneous language interactions between young children and caregivers, playmates, and teachers.[[1]](#footnote-1)(2) Using the Codes for Human Analysis of Transcripts (CHAT), a standardized, universally-used and multiple-level language coding system, to manually transcribe the language samples, and (3) using the Computerized Language ANalysis (CLAN) program to analyze the transcripts standardized by CHAT, researching various aspects of language usage, such as lexicon, syntax, morphology, phonology, discourse, and narrative.

There are two studies that have examined and compared the narrative abilities of autistic children and their TD peers using the *Tuesday* picture book and transcribed the recordings to text using the CHAT coding system (Rumpf, Kamp-Becker, Becker, & Kauschke, 2012; Kuijper, Hartman, Bogaerds-Hazenberg, & Hendriks, 2017). We briefly introduced the two studies in the following paragraphs:

First, (Rumpf, Kamp-Becker, Becker, & Kauschke, 2012) analyzed 31 8-12-year-old German-speaking children’s narratives in telling the story of *Tuesday*. Participants included 11 children with Asperger Syndrome (AS), nine children with Attention Deficit Hyperactivity Disorder (ADHD), and 11 Healthy Controls (HC). The analysis focused on the following linguistic categories: narrative length, sentence structure, sentence complexity, coherence and cohesion of stories, speaker's perspectives, and narrative styles. Researchers found that children with AS had limited abilities in several aspects of storytelling skills, especially in coherence of stories. They concluded that children with AS tended to tell shorter stories, use fewer pronoun references, and were less able to convey the main content of the story.

Second, (Kuijper, Hartman, Bogaerds-Hazenberg, & Hendriks, 2017) analyzed the verbal narratives of the story *Tuesday* from 106 6-12-year-old Dutch-speaking children, including 36 autistic children, 34 children with ADHD, and 36 TD peers. The researchers examined the following linguistic characteristics of the narratives: verbal productivity, speech fluency, syntactic complexity, lexical semantics, and discourse pragmatics. They found that both ASD and ADHD children exhibited impairments in telling stories compared to the TD peers, including difficulties in producing narratives coherently and cohesively, challenges in syntactic complexity, and excessive use of repetitions.

Previous studies also reported that autistic children show unique patterns in narrative productivity (i.e., story length or sentence length), pronoun use, syntactic skills, and personal style. However, these characteristics were reported inconsistently in different studies when using different measurements. The only consistent finding is that children with high-functioning ASD have the same lexical diversity as their age-matched TD peers. Other narrative characteristics of autistic children remain inconclusive. It is worth noting that the above study only used traditional statistical methods to conduct linguistic analysis of children's narratives. In this paper, we use an automated way to conduct statistical analysis of children's narratives and also use the results to improve the performance of our NLP model.

## Chinese Language Sample Analysis

To collect and analyze the child language data systematically, we use the Chinese Language Sample Analysis (CLSA) procedure to guide our data collection process (Tsai, 2009). CLSA was developed based on the analysis of English language samples initially and modified to fit the characteristics of Chinese. The CLSA procedure is primarily used to assess children's expressive linguistic abilities in natural contexts. According to CLSA, researchers collect children's spontaneous verbal language in natural settings through video and audio recordings. The language samples were then transcribed verbatim using specific transcription principles, such as CHAT. CLSA provides clear guidelines for sentence and word segmentation based on the characteristics of Mandarin Chinese. Researchers may later systematically analyze the transcripts based on the language characteristics representing different language developments. CLSA analysis can assess children's linguistic abilities, including phonology, vocabulary, grammar, and pragmatics. It also exhibits high sensitivity in detecting children’s improvement in language development (黃瑞珍、吳尚諭、蔡宜芳、黃慈芳、鄭子安, 2016). In this research, we adopted CLSA to guide the sample collection, transcription, and analysis processes.

The following are the CLSA guidelines for sample collection, transcription, and analysis of Chinese-speaking children's language samples.

**Sample Collection:**

1. Examiners who use CLSA should receive formal test administration training.
2. CLSA  only applies to children whose Mean Length of Utterance-characters (MLU-c) ≥ 5.
3. CLSA can only be used for children who have the ability to speak and without hearing or visual impairments.
4. Examiners need to collect children’s demographic information, including:
   1. Name
   2. Gender
   3. Chronological age
   4. Family situation:
      1. Home ranking
      2. Primary caregiver
      3. of primary caregiver’s education level
      4. The primary language used in the family
      5. The occupational name of the main source of income
      6. The child has the disability identified or not
      7. Contact information
      8. Other special matters
5. Video camera or recording equipment operates normally.
6. A well-lit and quiet environment
7. Only the examiner and the child are in the room
8. Interact with the child for a few minutes before formal recording to reduce fear, discomfort, or resistance.
9. There is no time limit for sample collection (in principle, each data collection should last 20-30 minutes)
10. Examiners encourage the child to talk by facilitating and supporting utterances when the child is unresponsive or responds infrequently.

**Transcription:**

1. Utterance segmentation rules:
   1. A child speaks a passage of speech apparently in one breath without a pause, often expressing a pause with the terminal intonation, called an utterance.
   2. If the pause time exceeds two seconds, it is an indicator of a clearly cut utterance.
   3. When children use "then," "and" and other related words to link sentences, each sentence should be divided into an independent utterance.
   4. When the pause, tone change, deep breath, etc. appear in the sentence, the sentence should be split.

2. Word segmentation rules:

Use Chinese Word Segmentation System developed by Academia Sinica (Academia Sinica, n.d.), and manually corrected the segmentation according to the word segmentation rules of the Taiwan Corpus of Child Mandarin (TCCM) (Cheung, Chang, Ko, & Tsay, 2011).

**Analysis:**

After the audio recordings are transcribed into text, the CHAT and CLAN will be used for language sample analysis. CLSA assesses several language characteristics to determine children's linguistic ability.

1. Mean length of utterance (MLU)

MLU is calculated by the total number of characters or words in the selected valid utterances divided by the total number of selected utterances. Autistic children often show differences in MLU compared to their typically developing peers.

1. Mean length of the five longest utterances (MLU5)

In addition to MLU, MLU5 can can offer supplementary insights into child language development. MLU5 is calculated by the total number of characters or words in the five longest sentences in the language sample, then divides the total number by five to take the average. In terms of Chinese, (周競、張鑑如, 2009) indicates that MLU5 can reflect the highest performance level of children in the complexity of sentences.

1. Vocabulary Diversity (VOCD)

Vocabulary diversity is often used to evaluate children's vocabulary ability in spontaneous language, and Type-Token Ratio (TTR) (Malvern & Richards, 1997) is a commonly used approach to analyze the vocabulary diversity.

TTR indicates the proportion of distinct words to the overall word count in a given language sample. Although TTR can reflect children's use of distinct words, it is more likely to be affected by the size of the total vocabulary. To address the TTR problem, (McKee, Malvern, & Richards, 2000) introduced the VOCD, a mathematical algorithm that incorporates TTR, as another supplemental indicator of vocavulary diversity. The first step of VOCD involves estimating the level of lexical diversity in a text. This is achieved by randomly selecting 100 samples, each consisting of 35 tokens, from the text and computing the average TTR for these samples. The same procedure is repeated for samples of 36 tokens, 37 tokens, and so on, up to samples of 50 tokens. Subsequently, the program generates a TTR curve for the text by plotting the average TTR value for each sample size, thus producing a random sampling of TTR values. After plotting the average TTR values for a random sample of 35-50 tokens, VOCD applies a formula called the D coefficient (Malvern D. , Richards, Chipere, & Durán, 2004) to generate a theoretical curve that closely matches the TTR curve of the random sample. Therefore, VOCD can minimize the impact of the number of sample words. The larger the VOCD value is, the higher the lexical diversity will be.

In Chinese, sentences are made up of words, and words are made up of characters. (吳啟誠, 2002) found that calculating MLU with words (MLU-w) as a unit has higher test-retest reliability than calculating MLU with characters (MLU-c) as a unit. Therefore, in our study we used MLU-c，MLU-w, MLU5-c, MLU5-w,VOCD-c, VOCD-w as measures for analysis.

1. Word analysis

According to grammatical functions, Chinese words can be divided into two categories: notional words and function words. Notional words are meaningful words, including noun, verb, adjective, numeral, measure word, pronoun, and adverb. Function words are words that can not stand alone as a meaningful segment but have grammatical meaning or carry out specific language functions,such as preposition, conjunction, auxiliary word,andinterjection. Word analysis can be used to assess children's vocabulary usage for each part of speech.

## Computational models for analyzing narration

Previous research on children’s narrative skills mainly uses manual analytic methods to investigate language characteristics across different developmental stages and populations with various language abilities. Only a few studies used NLP tools or neural networks to investigate the narrative performance differences between autistic individuals and TD peers.

To quantitatively characterize narrative performance in autistic individuals, sentiment and language abstraction analyses based on the Linguistic Category Model were employed (Chojnicka & Wawer, 2020). Their experiments included 50 Polish-speaking children (25 with ASD and 25 controls with TD, aged 7-25). Language samples were generated during two standardized tasks from the ADOS-2: Telling a Story from a Book and Description of a Picture. The results supported the sentiment and language abstraction analyses as a valuable tool. Specifically, autistic individuals demonstrated lower levels of language abstraction compared to TD individuals and exhibited difficulties expressing emotions.

In Chojnicka & Wawer’s work, they collected 50 Polish-speaking individual’s narratives in picture book (25 autstic participants and 25 TD controls, aged 7-25). The aim of the research was to identify the narratives produced by autistic individuals. Two text encoders, Embeddings from Language Models (ELMo) and Universal Sentence Encoder (USE), were utilized, along with three classification algorithms: XGBoost, support vector machines, and a dense neural network layer (Wawer & Chojnicka, 2022). Their experiments, employing deep neural network text representation models, demonstrated higher sensitivity, specificity, positive predictive values, and negative predictive values compared to human raters. However, these values were lower when compared to the two standardized instruments used in the study: ADOS-2 and Social Communication Questionnaire (SCQ) (Rutter, Bailey, & Lord, 2003). The SCQ consists of a series of 40 questions, which are answered by a parent, caregiver, or teacher who knows the individual well. The questions cover various aspects of social communication, such as social interaction, verbal and nonverbal communication, and restricted and repetitive behaviors commonly associated with autism. Each item in the SCQ is scored based on the presence or absence of specific behaviors related to ASD. The total score provides an indication of the likelihood that the individual may have autism, with higher scores suggesting a higher likelihood of the condition.

# Data Collection

图示

描述已自动生成

Figure : The flow chart of our experiment

As shown in Figure 1, the whole experiment can be divided into 5 steps: collect language sample by audio recording, transcribe the recording into text, verbal productivity and word analysis, feature extraction and classification, and acquire the prediction result. In this section, we introduced how we obtained recordings and transcribed them into text files. We also explained how we conduct verbal productivity & word analysis. In the next section, we described several techniques we applied to feature extraction and classification.

## Participants

A total of 23 school-age children were enrolled, including seven children diagnosed with ASD (ASD group) (five mild, one moderate and one severe) and 16 age-matched TD controls (TD group). The gender proportion of our participants reflected the fact that boys are more frequently diagnosed with ASD than girls (Table 1). In terms of family background, the educational background of the main caregivers is concentrated in the general/vocational high school and university (seven general/vocational high school, 13 university); The occupation distribution showed diversity, but in the ASD group, freelancing and housekeeping accounted for more than five, in the TD group, engineer accounted for the largest number of five.

|  |  |  |
| --- | --- | --- |
| Table : Group descriptive characteristics. | | |
| Group | Mean age (SD)  Range | Sex (M:F) |
| ASD (*n*=7) | 10.5 (0.85)  9.3-12 | 7:0 \*\* |
| TD (*n*=16) | 10.28 (1.37)  7.6-12 | 5:11 |
| SD: standard deviation; ASD: autism spectrum disorder; TD: typically developing.  \*\**p*<.01. | | |

## Procedure

The Central Regional Research Ethics Committee at China Medical University reviewed and approved the research procedure (IRB number: CRREC-112-002). Researchers contacted potential participants from the Hsinchu Association of Autism (社團法人新竹市自閉症協進會) and Hsinchu East District Zhulian Elementary School (新竹市東區竹蓮國民小學). The inclusion criteria for recruiting autistic children were: (1) age between 6 and 12 years, (2) using Mandarin Chinese as the primary language, (3) do not have hearing or visual impairments, (4) having the ability to speak, and (5) having received a medical diagnosis or disability identification as ASD. The exclusion criteria for the TD group were: (1) do not have a personal or family history of ASD, (2) do not have a history of developmental disorders, and (3) do not have neurological or psychiatric disorders or suspected genetic syndromes and developmental issues. All children in the ASD group received the ASD diagnosis within the validity period (ICD 9 code 299, ICD 10 code F84).

## Materials

In order to mitigate the limitations associated with a single picture book, we collected the language samples using two distinct picture books with varied styles and cultural backgrounds. In this study, we selected the following two picture books: *Tuesday* from the ADOS-2 module 3 and*子兒，吐吐 (Spit the seeds)* (李瑾倫, 1993). Below is the brief introduction of the two books:

*Tuesday* contains 29 pages and narrates the thrilling adventures of a colony of frogs.As shown in the left side of Figure 2, the drawing style is abstract, the author presents the storyline through the way of splitting mirrors. *Tuesday* does not include text in the book and lets the readers interpret the series of pictures freely.

Created by a Taiwanese writer, the story of*子兒，吐吐* is more situated in the local culture where the children live. The storyis about a piglet who eats papaya too fast and swallows the seeds. The piglet worries that the seeds will grow in his body and that he will become a papaya tree. As shown in the right side of Figure 2, the illustrator in*子兒，吐吐* used the hand-painting drawing style to give readers a light, vivid, and lovely artistic conception and make the book more accessible for children. Compared to 墙上挂着一幅画

低可信度描述已自动生成地图

低可信度描述已自动生成*Tuesday*, *子兒，吐吐*contains some Chinese text to describe the story.

Figure : Snapshots of the two selected picture books. Left side is from Tuesday, and Right side is from *子兒，吐吐*

## Get recording and transcription

According to CLSA's guidance, we recorded the child’s storytelling using iPhone 12 in a well-lit, quiet classroom with only the examiner and the child. Due to children's resistance or fear of cameras and parents' concerns about video recording, we only collect the language sample in audio recordings. Prior to the formal data collection, the examiner interacted with the tested child for two minutes to reduce their fear, discomfort, or resistance, establishing a shared focus of attention with the child. During the pre-test interaction, we invited the child to share their name, age, grade, favorite color and animal, and reasons. When the child started reading picture books, the examiner let the child choose which book to read first. When the child read the picture book of *Tuesday*, the examiner provided the following instructions: “It is a story about frogs. Can you read the book and tell me the story?” When the child started reading the picture book of *子兒，吐吐*, the examiner would ask the child if he or she understood the text in the picture book. If so, after reading the picture book, the examiner would ask the child to describe what happened to the Piglet. There was no time limit for recording, with a maximum of 20 minutes per picture book in principle. When the child did not respond or responded little, the examiner may provide more guided assistance to encourage the child to share more, such as “What is happening here?”

The transcriber independently transcribed all the collected language samples based on the CHAT transcription manual. The transcriber also made the segments of utterances, words, pauses, and repetitions based on the CLSA's guidance. A single utterance is defined as a child speaking a series of words or morphemes apparently in one breath without a pause. A sentence is considered as two utterances if there are more than two seconds of pause between two words. The transcriber used both words and characters to transcribe the language samples. In this study, we used the Chinese Word Segmentation System, an automatic Chinese word segmentation program developed by Academia Sinica (Academia Sinica, n.d.), to conduct the word segmentation. We also manually checked and corrected the segmentation results based on the word segmentation rules of the Taiwan Corpus of Child Mandarin (TCCM) (Cheung, Chang, Ko, & Tsay, 2011).

## Verbal productivity measures

We analyzed the transcribed texts based on the CLSA analysis guideline. We compared children's narrative skills between the ASD and TD groups. As shown in Table 2, we examined children’s verbal productivity by calculating the number of utterances, number of characters, and number of words. At the sentence level, we calculated the MLU and MLU5 in characters and words. We also measured the lexical diversity in children's narratives by calculating the VOCD in character and word. Finally, we investigated whether children in different groups have different preferences between the two picture books.

Table 2 shows the comparisons of verbal productivity between the ASD and TD groups. First, autistic children tended to narrative less than the TD children and have less numbers of utterances, characters, and words when telling stories. This is especially evident in the following five measures: MLU-c, MLU5-c, VOCD-c, MLU-w, MLU5-w. The independent samples t-test reveals that the ASD and TD groups differ significantly in (MLU-c: *p* = .03), (MLU5-c: *p*= .03), (VOCD-c: *p*= .037), (MLU-w: *p*= .025), and (MLU5-w: *p*= .016), while the differences between (VOCD-w: *p*= .065) and (book preference: *p*= .61) are not significant. The findings suggest that autistic children tended to use a limited range of characters and words when speaking. However, paradoxically, they show an adept use of lexical diversity.

|  |  |  |
| --- | --- | --- |
| Table : Narrative performance on verbal productivity measures | | |
| Verbal productivity measures | ASD | TD |
| Mean (SD) | Mean (SD) |
| Number of utterances | 69.29 (44.48) | 83.94 (25.67) |
| Number of characters | 452.43 (354.76) | 643 (235.68) |
| Number of words | 277.71 (220.28) | 409.13 (149.38) |
| MLU-c | 5.69 (2.57)\* | 7.47 (1.17) |
| MLU5-c | 12.11 (7.27)\* | 17.16 (3.27) |
| VOCD-c | 47.01 (29.43)\* | 66.92 (14) |
| MLU-w | 3.54 (1.62)\* | 4.75 (0.82) |
| MLU5-w | 7.43 (4.46)\* | 11.09 (2.29) |
| VOCD-w | 69.14 (58.05) | 115.06 (49.55) |
| Preference | 3:4 | 5:11 |
| SD: standard deviation; ASD: autism spectrum disorder; TD: typically developing; MLU-c: mean length of utterance-characters; MLU5-c: mean length of the five longest utterances-characters; VOCD-c: vocabulary diversity-characters; MLU-w: mean length of utterance-words; MLU5-w: mean length of the five longest utterances-words; VOCD-w: vocabulary diversity-words; Preference: Which book do children prefer? (*Tuesday*:*子兒，吐吐*).  \**p*< .05; \*\**p*< .01. | | |

## Word analysis

In addition to the macroscopic statistics of verbal productivity, we also conducted more detailed statistical analyses of words. We classified all transcribed words into two main categories: notional words (including nouns, verbs, adjectives, numerals, measure words, pronouns, and adverbs) and function words (includeing prepositions, conjunctions, auxiliary words,and interjections).

Table 3 shows the results of word analysis in notional and function words. Overall, the proportion of notional words used by autistic children is higher than that of TD children (notional word: *p*= .022), and the use of function words is lower than that of TD children (function word: *p*= .021). Autistic children are unable to fluently use measure words and auxiliary words compared to the TD children (measure word: *p*= .006) and (auxiliary word: *p*= .004). However, autistic children used adjectives (adjective: *p*= .036) more often than TD children in their narratives. This finding may be attributed to the prevalent utilization of narrative discourse among autistic children primarily for declarative statements or imperative expressions, resulting in limited usage of measure words and auxiliary words. Notably, this represents a distinctive pattern, as autistic children typically exhibit narrative expression deficits compared to TD children, while concurrently demonstrating relative proficiency in the specific domain of adjective usage.

|  |  |  |
| --- | --- | --- |
| Table : Word analysis | | |
| Parts of speech | ASD | TD |
| Mean (SD) | Mean (SD) |
| Noun | 34.52% (0.23) | 22.64% (0.06) |
| Verb | 19.80% (0.07) | 20.62% (0.03) |
| Adjective | 9% (0.09)\* | 4.04% (0.01) |
| Numeral | 2% (0.02) | 2% (0.01) |
| Measure word | 2% (0.02)\*\* | 3.93% (0.01) |
| Pronoun | 9% (0.07) | 11.35% (0.03) |
| Adverb | 9.62% (0.07) | 13.67% (0.03) |
| Notional word | 85% (0.1)\* | 78.65% (0.03) |
| Preposition | 3% (0.02) | 4.91% (0.02) |
| Conjunction | 7% (0.05) | 7% (0.04) |
| Auxiliary word | 4% (0.03)\*\* | 7.72% (0.02) |
| Interjection | 0% (0) | 1% (0.01) |
| Function word | 14% (0.1)\* | 20.82% (0.03) |
| SD: standard deviation; ASD: autism spectrum disorder; TD: typically developing.  \**p*< .05; \*\**p*< .01. | | |

# Method

In this section, we describe several techniques we applied to feature extraction & classification.

## Data augmentation

As we only recruited seven autistic children and 16 TD children, which is far from enough data for the subsequent training of the model. In order to increse the size of data, we used easy data augmentation techniques proposed by (Wei & Zou, 2019) to do data augmentation.

For a given sentence in the training set, we randomly chose and performed one of the operations in Table 4. To balance the amount of data between ASD and TD children, we amplified the data for ASD children by 5,10,15,20 times, and 2,4,6,8 times for TD children. Figure 3 shows the visualized result of the augmented ASD and TD datasets for picture book *Tuesday*. We applied Term Frequency-Inverse Document Frequency (TF-IDF) (Jones, 1972) to get feature vectors, k-means (MacQueen, 1967) for clustering, Principal Component Analysis (PCA) (Pearson, 1901) for dimension reduction, and Matplotlib[[2]](#footnote-2) to plot 2-D latent space representations.

As shown inFigure 3, we found that the resulting latent space representations for augmented sentences closely surrounded those of the original sentences, which were divided into 7 clusters (ASD dataset) and 16 clusters (TD dataset) in the figure. This showed that in most cases, augmented sentences conserved the labels of the original sentences.

|  |  |
| --- | --- |
| Table : Data augmentation operation | |
| Operation | Sentence |
| None | 有一個超級青蛙來到這裏，然後看到一個老奶奶在裏面，青蛙在看電視。 |
| There is a super frog that comes here and sees an old lady inside, and the frog is watching TV. |
| SR | 有一個超級青蛙來到**這邊**，然後看到一個**老婆婆**在裏面，青蛙在看電視。 |
| There is a super frog that comes **this side** and sees an **old granny** inside, and the frog is watching TV. |
| RI | 有一個超級青蛙來到這裏，然後看到一個老奶奶在裏面**坐著**，青蛙在看電視。 |
| There is a super frog that comes here and sees an old lady **sitting** inside, and the frog is watching TV. |
| RS | **超級青蛙有一個**來到這裏，然後看到**老奶奶**在裏面**一個**，青蛙在看電視。 |
| **super frog There is a** that comes here and sees **old lady** inside **an**, and the frog is watching TV. |
| RD | 有一個青蛙來到這裏，看到一個老奶奶在裏面，青蛙在看電視。 |
| There is a frog that comes here sees an old lady inside, and the frog is watching TV. |
| SR: synonym replacement; RI: random insertion; RS: random swap; RD: random deletion. | |

图表, 散点图

描述已自动生成 图表, 散点图

描述已自动生成

Figure : Visualization of augmented datasets for picture book *Tuesday*

## Neural networks in detecting ASD

In our study, we compared several methods for classifying participants with ASD and TD, using both established and novel techniques described in the literature. The methods we tested differed not only in the various ways they represented utterances, but also in how they handled verbal productivity and words analysis data or classification algorithms.

To evaluated the predictive ability of the model with limited data samples, we employed k-fold cross-validation. In each iteration, one of the k folds was used as the validation set, while the remaining k-1 folds were used as the training set. After each iteration, the performance metrics, such as accuracy, was recorded. The final performance of the model was typically reported as the average of these k iterations.

In our experiments we used 5-fold cross-validation.

### Support Vector Machines (SVM)

The first method we applied to represent utterances is TF-IDF (Jones, 1972). It is a widely used text feature representation method that evaluates the importance of a word in a document collection. It combines the concepts of term frequency (TF) and inverse document frequency (IDF) to measure the significance of a word within a document collection. TF refers to the frequency of a word appearing in a document. Generally, a higher term frequency indicates that the word is more important in the document. IDF represents the rarity of a word across the document collection. It is calculated by dividing the total number of documents by the number of documents containing the word and taking the logarithm of the result. The inverse document frequency can measure the uniqueness of a word. When a word appears in a majority of the documents, its IDF value will be lower, indicating lower importance. TF-IDF aids in identifying keywords, filtering out noise words, and calculating document similarity in information retrieval.

SVM (Pedregosa, et al., 2011) is a widely used machine learning algorithm for binary classification problems. It aims to find an optimal decision boundary, known as a hyperplane, that effectively separates two different classes of data. In SVM, the training samples are treated as data points in a feature space. By selecting a subset of training samples known as support vectors, SVM determines a decision boundary (hyperplane) that maximizes the margin between the two classes, which is referred to as the maximum margin classification. During prediction, SVM classifies new unseen samples into their respective classes based on the decision boundary. The support vectors surrounding the decision boundary determine the stability and generalization ability of the classification.   
 In our experiment setup, we initially employed data augmentation techniques to enhance the transcribed text dataset along with its corresponding labels. Subsequently, TF-IDF was applied to extract informative textual features from the augmented data. To classify the resulting feature vectors, we employed the SVM algorithm. Finally, a 5-fold cross-validation strategy was employed to assess the performance of the classification model.

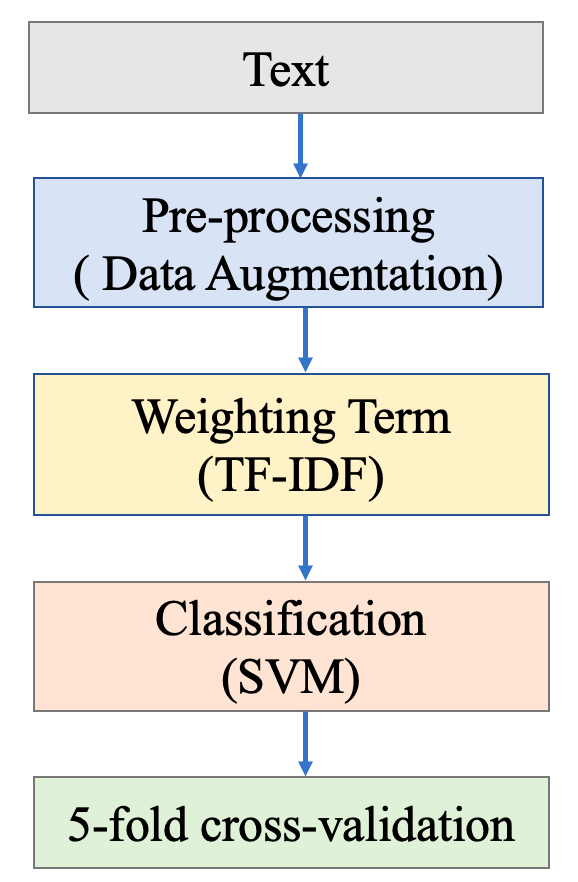


Figure : TF-IDF with SVM architecture

### Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin, Chang, Lee, & Toutanova, 2018) is a groundbreaking language representation model based on the Transformer architecture, which has made significant advancements in the field of NLP. BERT's key innovation lies in its ability to capture contextual information by considering both left and right contexts of words, unlike traditional language models that only focus on one direction. By utilizing the self-attention mechanism of Transformers, BERT achieves bidirectional encoding, enabling it to capture intricate semantic relationships and contextual information between words. The training process of BERT involves two stages: pretraining and fine-tuning. During pretraining, BERT leverages a large-scale unlabeled text corpus and learns universal language representations by performing tasks like masked language modeling and next sentence prediction. In the fine-tuning stage, BERT is further trained on specific tasks with labeled data, adapting the model to the requirements of the task at hand.

In our experiment, we used TinyBERT model (Jiao, et al., 2019). TinyBERT is a compact and efficient variant of BERT. It aims to compress the original BERT model while maintaining its performance and effectiveness. This is suitable for our small dataset mission, our data can be quickly trained in a lightweight network with a small number of parameters, and achieve good performance. In addition to simply using TinyBERT to train children's narratives, we also added verbal productivity and word analysis data as additional information to enhance the learning ability of the model. A simplified schema of encoding and classifying utterances using a transformer neural network such as TinyBERT is illustrated in Figure 5.

First, we amplified the transcribed documents and their corresponding labels, and then input the augmented data into TinyBERT of the 4-layer transformer architecture. At the same time, the corresponding verbal productivity and word analysis data were regularized and passed into the multilayer perceptron (MLP) with one linear layer. By concatenating the output of the MLP with the TinyBERT representation, we subsequently performed cross-attention and applied an additional layer of MLP for the final classification. Furthermore, we added the 5-fold cross-validation to assess the model performance.

图示

描述已自动生成

Figure : TinyBERT with hand coding and word analysis architecture

# Experiments

In this section, we discuss the data preprocessing performed on the datasets, the model configurations, and the details of four distinct experiments conducted.

## Data preprocessing

We divided the collected narratives (two different picture books) into six different datasets for the experiment.

* + Combine: Narratives of two picture books by 7 ASD and 16 TD participants. A total of 23 narrative documents.
  + Separate: Separate narratives of two picture books by 7 ASD and 16 TD participants. A total of 46 narrative documents.
  + Combine\_DA: Combine data with data augmentation.
  + Book 1\_DA: Picture book of *Tuesday* with data augmentation.
  + Book 2\_DA:Picture book of*子兒，吐吐*with data augmentation.
  + Separate\_DA: Separate data with data augmentation.

To make the model training more powerful, we adjusted the verbal productivity measures and words analysis data. We used StandardScaler from the Scikit-Learn package[[3]](#footnote-3) to standardize all features, so that the mean of the data was 0 and the variance was 1.

## Model settings

According to the suggestion of (Wei & Zou, 2019), we used α=0.05 (α is the proportion of changed words in each utterance) to do data augmentation. Since (Wei & Zou, 2019) focuses on enhancing English corpus data, we employed the Chinese stop word vocabulary[[4]](#footnote-4) developed by Harbin Institute of Technology and synonyms toolkit[[5]](#footnote-5) for the replacements.

We used the TinyBert\_4L\_zh[[6]](#footnote-6) as our pretrained model, which is a compact and efficient version of the BERT model specifically designed for Chinese language understanding tasks. The number of transformer layers *M*=4, the hidden size *d’*=312 and the head number *h*=12. For the training, we used 20 EPOCH, 5e-5 LR, 16 BATCH\_SIZE, and 200 Embedding\_dim.

## How much augmentation?

As mentioned in Section 4.1, we amplified the data for ASD children by 5, 10, 15, 20 times, and corresponding 2, 4, 6, 8 times for TD. In order to select the most appropriate amplification factor, we conducted experiments with the data of different amplification factors under the two architectures.

In the TF-IDF with SVM architecture, as illustrate in Figure 6 (a)-(d), the experiment results conducted on the four distinct datasets indicate that the accuracy exhibits an upward trend as the amplification factor increases. Notably, in the Book 1, Book 2, and Combine datasets the model achieves their peak performance when the ASD amplification factor is 15 times. Subsequently, the accuracy stabilizes or even declines. However, in the Separate dataset, the optimal performance is observed when the ASD amplification factor is increased by a factor of 20. A similar trend is observed in sensitivity (detection rate of children with ASD), wherein the Book 1, Book 2, and Separate datasets achieves their highest performance when the ASD amplification factor is 20 times. Conversely, in the Combine dataset, the best performance is attained when the ASD amplification factor is increased by a factor of 10. The specificity (the number of children detected as TD as a proportion of the total number of children with TD) is also on the rise, but unlike the accuracy and sensitivity, a smaller amplification factor can achieve higher performance.

With respect to the TinyBERT architecture, as illustrate in Figure 7 (a)-(d), the empirical findings derived from the four datasets demonstrate that the accuracy reaches its pinnacle when the amplification factor for ASD is set at 10 times. Subsequently, the accuracy is either stabilized or shows a decline. Similar patterns are observed in terms of sensitivity for the Book 2, Combine, and Separate datasets. However, the optimal performance in the Book 1 dataset is achieved when the ASD amplification factor is increased by a factor of 20. The distribution of the optimal specificity values varies across the datasets, although the overall discrepancy is not substantial.

Based on the above observations we found the most suitable data amplification factor for the two models: 15 times for SVM and 10 times for TinyBERT.

(a)(b) (c)(d)

Figure : Performance on SVM architecture with differentamplification factors

(a) (b)

(c)(d)

Figure : Performance on TinyBERT architecture with differentamplification factors.

## Comparison of the learning ability of autism characteristics between two models

Previous studies used the narrative of the picture book *Tuesday* for model training. In this experiment, in order to rule out the specificity caused by a single training set, we used two picture books as training sets to train the two models proposed previously. We used the Book 1 dataset as the training data, and the Book 2 dataset as the testing data. Then we used the Book 2 dataset as the training data, and the Book 1 dataset as the testing data. Through this experiment, we can see whether the two models can effectively learn the characteristics of autism, to do the classification.

Figure 8 presents an analysis of the two models’ performance on Book 1 for training and Book 2 for testing , with different data augmentation factors. The results indicate that the accuracy and sensitivity of the two models are very similar with data augmentation, showing an upward trend. However, the specificity shows large fluctuations when the amplification factor is less than 10. This is due to the fact that when the amount of data is too low, there are accidental phenomena and the model training is not sufficient.In general, TinyBERT has a relatively stable and better performance than TF-IDF with SVM. This is due to the fact that TinyBERT's language expression is very powerful and can capture semantic differences and word order well, even with a small amount of data.

(a) (b)

Figure : Two models’ performance on Book 1 for training and Book 2 for testing.

Figure 9 presents an analysis of the two models’ performance on Book 2 for training and Book 1 for testing with different data augmentation factors. The results indicate that the trends of the accuracy and sensitivity of the TF-IDF with SVM model are very similar with data augmentation. However, when the amplification factor is less than 10, the accuracy and sensitivity show a declining trend, and the specificity remains zero. This is likely due to the small data volume (for example, without data augmentation, we have only 23 narrative documents). However, when the amplification factor is more than 10, the performance measures show an upward trend. For the TinyBERT model, the trends of the accuracy, sensitivity and specificity are very similar with data augmentation. In general, TinyBERT has a relatively stable performance, and the three performance measures are similar with TF-IDF with SVM when the amplification factor is 20.

(a) (b)

Figure : Two models’ performance on Book 2 for training and Book 1 for testing.

From Figure 8 and Figure 9, it can be concluded that both TF-IDF with SVM and TinyBERT can learn certain autism characteristics to reach a sensitivity of 0.64-0.77 in the two testing sets. However, because the TF-IDF with SVM is easily limited by the amount of data, we use TinyBERT , and in the follow-up experiments.

## Is hand coding and word analysis helpful to the model?

Despite the remarkable achievement of a nearly 0.8 accuracy using the original data, there is still room for the improvement, especially for specificity. To address this problem, we leveraged the verbal productivity and words analysis data as supplementary information. This enhanced approach aims to augment the model's performance: especially in specificity. We used 4 datasets (Combine, Separate, Book 1 for training, and Book 2 for training dataset) that had not undergone data augmentation for experiments. As Table 5 shown, both architectures achieve the same accuracy for the Combine dataset, but the sensitivity and specificity are improved by 3% and 7%. In the Separate dataset, the accuracy is increased by 3%, the specificity is increased by 22%, and the sensitivity is decreased by 3%. In the task of selecting which book as the training set, the accuracy, sensitivity, and specificity are improved compared with pure TinyBERT. Especially when Book 2 is used as the training set, the sensitivity is increased by 5%, the specificity is increased by 13%, and the accuracy is increased by 9%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table : Verbal productivity and words analysis performance | | | | |
|  |  | Sensitivity | Specificity | Accuracy |
| TinyBERT | Combine | 0.76 | 0.4 | 0.78 |
| Separate | 0.86 | 0.51 | 0.8 |
| Book 1 for training | 0.82 | 0.71 | 0.75 |
| Book 2 for training | 0.79 | 0.79 | 0.76 |
| TinyBERT with hand coding and word analysis | Combine | 0.79 | 0.47 | 0.78 |
| Separate | 0.83 | 0.73 | 0.83 |
| Book 1 for training | 0.82 | 0.84 | 0.81 |
| Book 2 for training | 0.84 | 0.92 | 0.85 |
|  | | | | |

## Comparison with previous study

As mentioned in Section 2.4, (Wawer & Chojnicka, 2022) also used the *Tuesday* picture book to conduct experiments. They used ELMo, USE and three classification algorithms to detected narratives produced by autistic children, with 25 autistic individuals and 25 controls. We augmented the original data to the same number for comparison. In Table 6, we compare our three different methods with (Wawer & Chojnicka, 2022). The three architectures demonstrate evident enhancements in terms of sensitivity, specificity, and accuracy. Particularly, the TFIDF-SVM approach exhibites a remarkable improvement of 7%-8% in both sensitivity and accuracy, along with a substantial 16% increase in specificity comparing to (Wawer & Chojnicka, 2022). Furthermore, TinyBERT outperformes (Wawer & Chojnicka, 2022) by 14%-21% across all three performance metrics. Notably, the integration of verbal productivity and word analysis into TinyBERT yields exceptional results, surpassing the 90% threshold for sensitivity, specificity, and accuracy. These findings reaffirm the efficacy of incorporating supplementary knowledge through verbal productivity and word analysis, facilitating superior learning within the model.

|  |  |  |  |
| --- | --- | --- | --- |
| Table : Our methods compare with (Wawer & Chojnicka, 2022) | | | |
|  | Sensitivity | Specificity | Accuracy |
| (Wawer & Chojnicka, 2022) | 0.72 | 0.68 | 0.7 |
| TFIDF-SVM | 0.79 | 0.84 | 0.78 |
| TinyBERT | 0.86 | 0.89 | 0.84 |
| TinyBERT with hand coding and word analysis | **0.93** | **0.91** | **0.92** |
|  | | | |

# Conclusion

Our paper aimed to use NLP approach and language ability analysis from children's picture book narratives for autism tendency analysis. Thanks to the continuous progress of NLP research in recent years, we can use higher-performing technology to solve the problem. It is important to acknowledge that, thus far, the majority of research on ASD detection through speech has relied on assessment sheets or rudimentary language analysis tools, with only a limited number of researchers exploring computer-based methods. However, our study introduces three novel NLP architectures that exhibit enhanced precision and efficiency in ASD detection. These advancements hold promise for facilitating more accessible and expedient services in autism tendency analysis in the future.

For our proposed NLP method, the three architectures showed better performance in different tasks. First, we conducted experiments on TF-IDF with SVM architecture and TinyBERT under different data amplification factors. The experiment results showed that when the ASD amplification factor was 15 and 10 respectively, the two models had achieved relatively stable performance. The accuracy, sensitivity and specificity can reach more than 90%, among which TinyBERT can reach 99% and above.

In the second experiment, we compared two models’ learning ability of autism characteristics. We alternately used two different picture books as the training set in the experiments. The experiment results showed that both the TF-IDF with SVM and TinyBERT can learn certain autism characteristics, and reached a sensitivity of 0.64-0.77 in the two testing sets. TinyBERT was more stable than the TF-IDF with SVM which was easily limited by the amount of data.

In the third experiment, we examined whether verbal productivity and word analysis can improve the performance of the model. In this experiment, we did not use the dataset that had undergone data augmentation, but only used the collected original data. We conducted ablation experiments on 4 different datasets (Combine, Separate, Book 1 for training, Book 2 for training) and the results showed that verbal productivity and word analysis can improve the accuracy, sensitivity and specificity to a certain extent, in the case of a very small amount of data, which also proved the importance of the use of additional knowledge to improve the performance of the model.

In the last experiment, we compared the three architectures we proposed with the previous study. In order to have the same amount of data, we amplified the original data to be the same amount as the previous study (25ASD, 25TD). It can be seen from Table 6 that the three architectures we proposed surpassed (Wawer & Chojnicka, 2022). Among the three architectures, TinyBERT with verbal productivity and word analysis achieved the best performance, with the sensitivity, specificity, and accuracy all exceeding 90%.

Generally, the best overall results were achieved by the TinyBERT with verbal productivity and word analysis. In this setup, such supplementary knowledge leads to better classification performance. In addition to the models' higher performance, we also explored children's language narrative features. As previous studies have found, ASD children produce fewer words and utterances than TD children, and also weaker in language diversity. However, in our statistics, we were surprised to find that autistic children have a strong ability to use adjectives, which was not noticed in previous studies.

# Limitations and future work

It is undeniable that there are many limitations in our research. Primarily, it is important to note that the available dataset contains a limited number of samples, exclusively comprising males within the collected ASD group. Moreover, the gender distribution within the TD dataset is imbalanced. This inherent imbalance has the potential to introduce bias during subsequent analyses of language characteristics. Additionally, owing to the challenges associate with collecting personal data, we are unable to obtain information pertaining to the verbal IQ and non-verbal IQ of the participants. Consequently, the impact of IQ variability on the results can not be determined in this study. In addition, it is noteworthy that our experiment solely relies on audio recordings, without accompanying video recordings that can capture children's nonverbal and behavioral expressions. As a result, the absence of visual records limits our ability to analyze the nuances of children's nonverbal cues and gestures. Additionally, we utilize speech-to-text transcription for subsequent analysis, which hinders the inclusion of children's voice intonation in the recorded data. It is worth mentioning that achieving high-quality speech-to-text transcription poses a significant challenge. In our research, manual transcription is employed, which introduces the possibility of transcription errors and potential deviations during the transcription process.

Autism tendency analysis through children's narratives remains a difficult challenge, with many factors to consider. Future endeavors should incorporate additional language assessment methods, such as the analyses of sentence structure composition, narrative coherence, pitch variations, and speech fluency. In order to achieve a more nuanced understanding of the language characteristics specific to children with ASD, it may be advantageous to include a comparative analysis involving a cohort of language-impaired children. This comparative approach can help delineate the distinctive features associated with autism more effectively.

# Reference

Academia Sinica. (n.d.). *The Chinese word segmentation system*. Retrieved from The Chinese word segmentation system: http://ckipsvr.iis.sinica.edu.tw/

Akdemir, A., & Shibuya, T. (2020). Analyzing the Effect of Multi-task Learning for Biomedical Named Entity Recognition. *arXiv preprint*, p. arXiv:2011.00425.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (5th ed.,text rev.).*

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*, pp. 59, 234-254.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, pp. 21(1),37-46.

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*, pp. 28,193-204.

Chai, Z., Jin, H., Shi, S., Zhan, S., Zhuo, L., & Yang, Y. (2022). Hierarchical shared transfer learning for biomedical named entity recognition. *BMC bioinformatics*, pp. 23(1), 1-14.

CheungH., ChangJ.C., KoW.H., & TsayJ. (2011). Taiwan Corpus of Child Mandarin (TCCM). 107-110.

Chojnicka, I., & Wawer, A. (2020). Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. *PLoS One*, pp. 15(3),e0229985.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 51-60.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 51-60.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, pp. 10, 1178222618792860.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *In Proceedings of the international AAAI conference on web and social media*, (pp. 128-137).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: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*, pp. 34,83-98.

Engberg-Pedersen, E., & Christensen, R. (2016). Mental states and activities in Danish narratives: children with autism and children with language impairment. *Journal of Child Language*, 44(5),1192-1217.

Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, pp. 11(1), 1-17.

Goldstein, S., & Ozonoff, S. (2018). *Assessment of autism spectrum disorder.* Guilford Publications.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., & ... Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351.*

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, pp. 28(1), 11-21.

Khan, M. R., Ziyadi, M., & AbdelHady, M. (2020). Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *arXiv preprint*, p. arXiv:2001.08904.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations.*

Kuijper, S. J., Hartman, C. A., Bogaerds-Hazenberg, S., & Hendriks, P. (2017). Narrative production in children with autism spectrum disorder (ASD) and children with attention-deficit/hyperactivity disorder (ADHD): Similarities and differences. *Journal of Abnormal Psychology*, 126(1),63.

Li, J., Zhang, Z., Li, X., & Chen, H. (2008). Kernel‐based learning for biomedical relation extraction. *Journal of the American Society for Information Science and Technology*, pp. 59(5), 756-769.

Liao, Z., & Wu, H. (2012). Biomedical named entity recognition based on skip-chain Crfs. *In 2012 international conference on industrial control and electronics engineering* *(IEEE)*, pp. 1495-1498.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv*, p. preprint arXiv:1907.11692.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule–2nd edition (ADOS-2).* Los Angeles, CA: Western Psychological Corporation.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. . In *Probability* (pp. 281-297). Los Angeles LA USA: University of California.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of child language*, 17(2), 457-472.

Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., & ... & Shaw, K. A. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020. *MMWR Surveillance Summaries*, pp. 72(2),1.

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*, pp. 8(8),987-996.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, pp. 12,58-71.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). In *Lexical diversity and language development* (pp. 16-30). New York: Palgrave Macmilla.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing*, pp. 15(3),323-338.

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., & Jensen, L. J. (2013). The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PloS one*, pp. 8(6), e65390.

Parsons, L., Cordier, R., Munro, N., Joosten, A., & Speyer, R. (2017, December). A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PloS one*, 12(4), e0172242.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. . *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, pp. 2(11), 559-572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. . *The Journal of machine Learning research*, 12, 2825-2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.

Rumpf, A. L., Kamp-Becker, I., Becker, K., & Kauschke, C. (2012). Narrative competence and internal state language of children with Asperger Syndrome and ADHD. *Research in Developmental Disabilities*, 33(5),1395-1407.

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire.* Los Angeles: Western: Psychological Services.

Sah, W. H., & 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.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51,1928-1941.

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv*, p. preprint cs/0306050.

Schaaf, C. P., Betancur, C., Yuen, R. K., Parr, J. R., Skuse, D. H., Gallagher, L., & ... & Vorstman, J. A. (2020). A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nature Reviews Genetics*, pp. 21(6),367-376.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. *Algorithms*, pp. 13(1), 7.

Tong, Y., Chen, Y., & Shi, X. (2021). A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (pp. 4804-4813).

Tsai, I.-F. (2009). A Study of Chinese Language Sample Analysis for 3-5 Years Old Children.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin. (2017). Attention is all you need. *Advances in neural information processing systems*, (p. 30).

Wawer, A., & Chojnicka, I. (2022). Detecting autism from picture book narratives using deep neural utterance embeddings. *International Journal of Language &Communication Disorders*, pp. 57, 948–962.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196.*

Wiesner, D. (1991). *Tuesday.* New York, NY: Clarion Books.

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., & ...&Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Research*, pp. 15(5),778-790.

黃瑞珍、吳尚諭、蔡宜芳、黃慈芳、鄭子安. (2016). 華語兒童語言樣本分析. 新北市: 心理.

教育部 (2014). *身心障礙及資賦優異學生鑑定辦法說明手冊.*

李瑾倫 (1993). 子兒，吐吐. 臺北市: 信誼.

李宜融 (2015). 自閉症類疾患兒童的診斷穩定度、自閉症狀變化與認知及適應功能關係之縱貫研究. 擷取自 https://nccur.lib.nccu.edu.tw/bitstream/140.119/76221/1/019101.pdf

衛生福利部 (2023). *2023最新自閉症人數比例*.財團法人中華民國自閉症基金會: https://www.fact.org.tw/contents/news\_ct?id=149

吳啟誠 (2002). 語障兒童口語能力指標之信度探究（未出版之碩士論文）.嘉義縣, 國立嘉義大學.

周競、張鑑如 (2009). 漢語兒童語言發展研究：國際兒童語料庫研究方法的應用. 中國北京市: 教育科學出版社.

1. https://childes.talkbank.org/ [↑](#footnote-ref-1)
2. https://matplotlib.org/ [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [↑](#footnote-ref-3)
4. https://github.com/goto456/stopwords/blob/master/hit\_stopwords.txt [↑](#footnote-ref-4)
5. https://github.com/chatopera/Synonyms [↑](#footnote-ref-5)
6. https://github.com/chakki-works/seqeval [↑](#footnote-ref-6)