Applying machine learning to language problem analysis

Kuo-Chung Chu
Department of Information Management
National Taipei University of Nursing and Health Sciences
Department of Education and Research, Taipei City Hospital,
Taipei City, Taiwan (R.O.C.)
kcchu8992@gmail.com

Yu-Jen Chiu
Department of Information Management
National Taipei University of Nursing and Health Sciences
Taipei City, Taiwan (R.O.C.)
worklight107@gmail.com

Jakir Hossain Bhuiyan Masud
Public health informatics foundation, Dhaka, Bangladesh
jakir_msd@yahoo.com

***Abstract*—During childhood, language development plays a pivotal role, serving as the cornerstone for early learning and social integration while significantly influencing future academic and social accomplishments. Therefore, effective assessment and support of children's language development are paramount. Among the various assessment tools, language sample analysis stands out as a key method. This involves experts collecting, transcribing, and analyzing children's language samples to evaluate their language development. The present study endeavors to merge language sample analysis with artificial intelligence (AI) machine learning technology to establish a classification model capable of discerning between children with normal development and those potentially facing language-related challenges. To achieve this goal, the study leverages the ASDBank and CHILDES as data sources, representing children with potential language issues and those with typical development, respectively. Through a series of processes, the corpus undergoes language sample analysis, yielding eight characteristic indicators (MLU-w, MLU-c, MLU5-w, MLU5-c, CTTR-w, CTTR-c, VOCD-D, VOCD-c). Utilizing these indicators as model features, synthetic data technology is employed to address data insufficiency, resulting in the generation of six combined datasets, totaling 1,000 data points, with 854 entries of 91.50% quality. Subsequently, a supervised machine learning classification model is constructed utilizing Bayesian classification, random forest, and gradient boosting machine methods. Considering the appropriate quantity of synthetic data generated and overall model evaluation performance, the random forest model constructed emerges as the optimal classification model, yielding highest accuracy (0.76), sensitivity (0.94), and F1-score (0.85). The study's finding is that the appropriate quantity of synthetic data positively impacts model performance, with its effect stabilizing after reaching a certain threshold.**

***Keywords*—language sample analysis, language development issues, machine learning, synthetic data, binary classification**

I. INTRODUCTION

The issue of children's language development is of paramount importance. Acquiring the ability to effectively use speech and language for communication is a fundamental milestone in a child's development, with profound implications for their academic and social trajectories throughout life [1]. Children experiencing challenges in language development may exhibit a range of difficulties, varying in severity, including disorders affecting both language comprehension and expression. Among these challenges, autism spectrum disorder stands out as a prevalent condition characterized by impairments in language and speech development, significantly impacting interpersonal relationships. Early intervention has been shown to yield enduring positive outcomes for individuals diagnosed with autism [2]. Detecting potential language issues in children early on and promptly seeking professional evaluation and diagnosis can significantly enhance the effectiveness of early intervention strategies. In tracing the evolution of methods for assessing children's language development, two primary approaches have emerged: standardized testing and criterion-referenced assessment. Standardized tests involve comparing a child's performance against established norms to ascertain any significant deviations. Conversely, criterion-referenced assessment focuses on monitoring the progress of language development within the context of specific language disorder cases, without reference to established norms [3]. Recognizing the distinct advantages and limitations of both standardized and criterion-referenced tests, scholars have sought to introduce a third assessment method known as Language Sample Analysis (LSA). LSA entails the collection of natural language samples from children in everyday contexts, such as conversations or narratives, which are then transcribed and analyzed. This method offers a comprehensive evaluation of various aspects of language, including form, content, and usage, thereby providing sensitive and reliable measures of language development [4]. Notably, LSA complements standardized testing by focusing on spontaneous language use, addressing some of the limitations inherent in standardized assessments [5]. In practical application, speech therapists employ various assessment methods tailored to specific circumstances. They carefully select appropriate indicators to evaluate a child's oral expression abilities, enabling them to design tailored treatment plans. Consequently, therapists often integrate multiple assessment approaches, including the combination of traditional methods with LSA, which has proven to be a viable strategy.

The LSA evolution traces back to the realm of language acquisition research and public children's corpora. Initially, scholars like Chomsky posited that language is innate, with development linked to individual physiological processes. However, insights from cognitive scientists and psychologists such as Bates and Elman have underscored the role of learning in language acquisition. Consequently, extensive research focuses on recording natural language usage by children, converting audio or video files into text data, and conducting rigorous statistical analyses. Yet, this transcription process is notably time-consuming and labor-intensive, often requiring ten to fourteen hours to capture nuances like tone, expression, and coherence. Historically, the fragmented and undisclosed nature of language corpora collection led to inefficiencies and non-reusability [6-13]. Recognizing the importance of addressing challenges in collecting, accessing, and analyzing children's corpora, initiatives supported by organizations like the MacArthur Foundation, the National Institutes of Health, and the National Science Foundation (NSF) have emerged.

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00050

Among these, the CHILDES system (Child Language Data Exchange System) developed by Catherine Snow and Brian MacWhinney has significantly contributed to the standardization and exchange of child language sample data. Concurrently, within the domain of speech therapy, advancements in speech recognition technologies offer promising applications. Speech recognition tools are already assisting patients and healthcare professionals, poised to play an increasingly integral role in early disease detection [14]. Despite these advancements, the utilization of artificial intelligence (AI) data analysis methods as adjuncts for detecting suspected language disorders remains relatively underexplored.

Several application studies have explored the AI integration with language assessment, primarily relying on standardized assessment scores to construct models [15, 16]. However, there is a notable gap in research regarding the LSA utilization indicators for model development. Hence, this study aims to investigate the feasibility of building an effective AI model using language sample indicators as features. Combining children's language sample data with AI technology forms the basis of this research. The methodology involves analyzing publicly available language samples on the Internet and employing AI methods to assess the language proficiency of the children within the dataset. The primary objective is to utilize an AI classification model to identify children suspected of having language problems based on their language samples. LSA serves as a fundamental tool in the arsenal of speech therapists, involving the meticulous collection, processing, and analysis of language samples. However, this process is time and labor-intensive, particularly when dealing with a large volume of samples simultaneously. This often presents challenges for therapists in prioritizing evaluation tasks, leading to inefficiencies. Recognizing these challenges, this study endeavors to leverage AI technology to streamline the identification of children who may require immediate attention from speech therapists. By employing AI to classify children with potential language problems, therapists can efficiently prioritize evaluation tasks, ensuring timely intervention. If AI can successfully distinguish between children with and without language issues, it can provide valuable guidance to therapists regarding which language samples warrant priority evaluation.

II. Method

A. Materials

Utilizing publicly available children's corpora not only maximizes the value of vast datasets previously collected and transcribed by researchers but also alleviates the time and resource burden associated with repetitive data collection and transcription. Moreover, it ensures the continuity of past research endeavors. Consequently, this study opted to leverage the public children's corpora accessible through the TalkBank platform as a rich source of language development data. TalkBank hosts numerous sub-projects, each offering diverse language development corpora. For this study, data sourced from two specific sub-projects within TalkBank were utilized: CHILDES from the Child Language Banks sub-project and ASDBank from the Clinical Banks sub-project. ASDBank serves as a repository of language samples from children with language disorders [17], whereas CHILDES provides data from typically developing children [18].

B. Data pre-processing

Before proceeding with data preprocessing, it is necessary to download the original CHAT files of the selected corpus, namely the ASDBank Mandarin Shanghai Corpus [17] and the CHILDES Mandarin Chang Toy Play Corpus [18]. The data preprocessing phase consists of two main components: language sample indicator data extraction and generation, and data merging.

1. Language Sample Indicator Data Extraction:
   - Exclude corpus files lacking age recordings.
   - Prepare for subsequent analysis of word-based" indicators.
   - Utilize the CLAN program to analyze children's language indicator data.
   - Compile the language indicator data from both types of corpora and incorporate the Bollinger field data.

2. The process of generating and merging data involves two steps:
   - Insufficient Data Processing using Generative Adversarial Networks (GAN): This study employs Synthetic Data Vault (SDV) synthetic data technology developed by MIT to address insufficient research data.
   - Merge Original Data and Synthetic Data: Merge 146 original data with synthetic data generated by the CTGAN model.

Synthetic data, artificially generated information, serves as a substitute for real historical data and is crucial when the real dataset lacks sufficient quality, quantity, or diversity. The CTGAN model, within the SDV framework, is utilized for data generation. CTGAN is renowned for its ability to generate high-quality tabular profiles and has significantly contributed to synthetic profile generation research [19].

In previous studies, statistical analysis methods were employed to preliminarily identify children suspected of having language problems. However, with AI advancements, machine learning techniques have emerged as powerful tools for data analysis. This study focuses on classification problems and utilizes machine learning technology to classify data.

C. Metrices

Recent studies have broadened the LSA application as a research methodology. Particularly in the analysis of Chinese language samples, it has been observed that a higher total number of sentences analyzed in a case results in closer sample analysis index results among the tested preschool children. Certain indicators derived from child language sample analysis have demonstrated reliability and validity in distinguishing between typical children and those with language problems. The selected language sample analysis indicators for this study are presented in Table 1:

TABLE 1: SELECTED LANGUAGE SAMPLE ANALYSIS INDICATORS

Indicator	Explanation
MLU-w	average number of sentence words
MLU-c	average sentence word count
MLU5-w	average number of words in the longest five sentences
MLU5-c	average word count of the five longest sentences
CTTR-w	Corrected dissimilar word ratio
CTTR-c	Corrected dissimilar word ratio
VOCD, D	Vocabulary Diversity

VOCD, D-c	Diversity of Words

Verification of the machine learning model entails dividing the dataset into a training set and a test set for validation. In this study, the dataset is divided into 80% for training and 20% for testing. Given the study's focus on identifying data indicative of potential language problems, instances with language problems are labeled as True, while those without are labeled as False.

III. Results

Synthetic data, also known as artificial data, is a type of non-real data generated through artificial algorithms or simulation processes. It holds particular significance in machine learning and data science as it serves to train models effectively. In this study, the CTGAN Python library within the SDV framework is utilized to create and generate synthetic tabular data. Among the various parameters within the SDV/CTGAN data synthesis library, the epoch parameter stands out as particularly crucial as it directly influences the quality of the synthesized data. The quality of synthetic data is paramount as it directly impacts the accuracy of models generated through machine learning training data. Adjusting the epoch parameter allows for the generation of synthetic data of varying quality. Synthetic data can be evaluated and compared to real data to ensure similarity in statistical and mathematical properties. The evaluate_quality function is employed to assess the quality of synthetic data based on field shapes and correlations, facilitating comparison with real data. To assess the performance of the binary classification prediction model for children's language development, various machine learning evaluation indicators and prediction models are calculated, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These metrics are organized into a confusion matrix. Evaluation indicators cover accuracy (Acc), sensitivity (Sen), prevalence (Prev), specificity (Spec), precision (Prec), F1 score (F1-score), ROC curve (Receiver Operating Characteristic curve), and AUC (Area Under Receiver Operator Characteristic curve). The models of the machine learning models, namely Naive Bayes classifier (NB), Random Forest (RF), and Gradient Boosting Decision Trees (GBDT), are employed to calculate various performance indicators. In this study, synthetic data of varying numbers and qualities are first generated. Subsequently, 80% of the combined data is allocated as a training set, with the remaining 20% serving as the test set. Applied in the medical field, individuals with language problems are designated as positive examples (Positive), while those without language problems are labeled as negative examples (Negative). The research experiment systematically compares different scenarios of synthetic data generation and classification model combinations. A total of 1,000 data points were merged for analysis, comprising 854 synthetic data entries (epoch=10,000/quality=91.50%) and 146 original data entries, the performance indicators of each classification model were evaluated, as presented in Table 2, Under the conditions of a total of 1,000 merged data entries and a synthetic data quality of 91.50%, the classification models exhibit notable sensitivity (Sen) exceeding 0.8. Precision (Prec), representing the positive predictive value (PPV), reaches close to 80% when employing the Bayesian classifier (NB). Moreover, employing Combination 1 with the Bayesian classifier (NB), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT), indicating favorable model performance.

Table 2: Performance Indicators of classification models

Data quality 91.50%	Acc	Prec	Sen	Prev	Spec	F1	AUC
NB	0.71	0.79	0.82	0.75	0.36	0.81	0.66
RF	0.76	0.78	0.94	0.75	0.20	0.85	0.67
GBDT	0.71	0.76	0.88	0.75	0.20	0.82	0.67

IV. Conclusion

LSA serves as a vital assessment method in the toolkit of speech therapists, albeit demanding considerable manual effort from sample processing to analysis. In scenarios where therapists confront a large volume of language samples, prioritizing processing and evaluation becomes challenging. Introducing AI classification models capable of categorizing samples as either without language problems or with language problems could aid therapists in decision-making regarding treatment and assessment prioritization for children. However, it's crucial to acknowledge the limitations of this study, which solely focuses on analyzing experimental model data derived from its dataset. Whether such a hypothesis can be effectively applied in clinical settings requires further development and verification through additional research. This study initiates from a dataset of children's language samples publicly available on the internet, albeit with limited data sources. Consequently, prior to constructing classification models, AI synthetic data technology, specifically the SDV CTGAN model within the Python library, was employed to address data insufficiency. This function library facilitates the generation of synthetic data and aids in verifying its quality. The study amalgamated the original data with six sets of synthetic data varying in quality and quantity. Subsequently, the NB, RF, and GBDT methods in AI were employed to establish classification models. These trained models can effectively categorize corpus sources into those without language problems and those with language problems, offering valuable insights for future research endeavors. It's important to note that clinical diagnosis by a speech therapist is a nuanced and professional process, and this study does not serve as a diagnostic basis. Typically, therapists determine whether a child undergoing testing requires a language sample assessment and whether it should be supplemented with other assessment methods for a comprehensive diagnosis. Therefore, the contribution of this study lies in its ability to provide a reference for future research by analyzing AI model data.

Acknowledgment

This research is partially supported by the National Science and Technology Council, Taiwan (NSTC 112-2637-E-227-001).

References

[1] 1)A. P. Kaiser and M. Y. Roberts, "Advances in Early Communication and Language Intervention," Journal of Early Intervention, vol. 33, no. 4, pp. 298-309, 2011/12/01 2011, doi: 10.1177/1053815111429968.

[2] 7) J. L. Matson, J. Wilkins, and M. González, "Early identification and diagnosis in autism spectrum disorders in young children and infants: How early is too early?," Research in Autism Spectrum Disorders, vol. 2, no. 1, pp. 75-84, 2008.

[3] 14)R. Paul, Language disorders from infancy through adolescence: Assessment & intervention. Elsevier Health Sciences, 2007.

[4] 19)B. Leadholm and J. Miller, "Language sample analysis: The Wisconsin guide. Madison: Wisconsin Department of Public Instruction; 1992," ed: Analysis.

[5] 20)M. N. Hegde and C. A. Maul, Language disorders in children: An evidence-based approach to assessment and treatment. Pearson College Division, 2006.

[6] 23)N. A. Chomsky, Reflections On Language. Temple Smith, 1975.

[7] 24)N. Chomsky, Language and Problems of Knowledge: The Managua Lectures (no. 2). MIT Press, 1987, pp. 5-33.

[8] 25)M. Piattelli-Palmarini, "Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language," (in eng), Cognition, vol. 31, no. 1, pp. 1-44, Feb 1989, doi: 10.1016/0010-0277(89)90016-4.

[9] 26)S. Pinker, The Language Instinct: How the Mind Creates Language. 1994.

[10] 27)E. Bates, "On the nature and nurture of language," Frontiere della biologia il cervello di homo sapiens [Frontiers of biology: the brain of homo sapiens]. Instituto della Enciclopedia Italiana, pp. 241-65, 1999.

[11] 28)E. Bates and J. Elman, "Learning rediscovered," Science, vol. 274, no. 5294, pp. 1849-1850, 1996.

[12] 29)E. Bates and G. F. Carnevale, "New directions in research on language development," Developmental review, vol. 13, no. 4, pp. 436-470, 1993.

[13] 30)P. Fletcher and B. MacWhinney, The handbook of child language. Blackwell Oxford, 1995.

[14] 11)J. L. a. V. Berisha. "How Will Artificial Intelligence Reshape Speech-Language Pathology Services and Practice in the Future?" ASHA JOURNALS ACADEMY. https://academy.pubs.asha.org/2020/08/how-will-artificial-intelligence-reshape-speech-language-pathology-services-and-practice-in-the-future/ (accessed.

[15] 12)L. Gasparini et al., "Using machine-learning methods to identify early-life predictors of 11-year language outcome," Journal of child psychology and psychiatry, and allied disciplines, 2022, doi: 10.1111/jcpp.13733.

[16] 13)A. Borovsky, D. Thal, and L. B. Leonard, "Moving towards accurate and early prediction of language delay with network science and machine learning approaches," Sci Rep, vol. 11, no. 1, p. 8136, Apr 14 2021, doi: 10.1038/s41598-021-85982-0.

[17] 46)ASDBank Mandarin Shanghai Corpus. [Online]. Available: https://doi.org/10.21415/T5HW46

[18] 47) CJ Chang. CHILDES Mandarin Chang Toy Play Corpus. [Online]. Available: https://childes.talkbank.org/access/Chinese/Mandarin/ChangPlay.html

[19] 53)L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," Advances in neural information processing systems, vol. 32, 2019.