“Use of NLP and ML techniques to clinically assess children Autism Spectrum Disorder and Specific Language Impairment”

The Prediction Problem

Research Question Formulation:

The traditional approach to diagnosing Autism Spectrum Disorder (ASD) has faced significant challenges, including inconvenience and potential inaccuracy. The traditional method typically required the presence of clinicians and trained linguists. Additionally, diagnoses were often based on comparisons with the average scores of a reference population. Children scoring at least 1.25 times below the standard deviation (SD) in two or more measures were considered to have Language Impairment (LI). However, Campbell et al. (1997) noted that such norm-referenced tests could be biased, leading to inaccuracies (Lee, 2016, pg. 4).

To address these issues of inaccessibility and inaccuracy, the research from Lee has proposed the use of Natural Language Processing (NLP) and Machine Learning (ML). This approach involves using NLP to extract linguistic features indicative of Autism Spectrum Disorder (ASD) and Specific Language Impairment (SLI). Machine Learning algorithms would then be employed to predict the likelihood of a child having LI (ASD or SLI) based solely on text analysis (Lee, 2016, pg. 1). Specifically, using classification, the expected result after the ML prediction will be “predicting a label for the data set with an acceptable accuracy rating” (Lee, 2016, pg. 8). This technological advancement aims to improve the diagnostic process for ASD, making it more accessible and accurate (Lee, 2016, pg. 4). However, indeed being a good approach, having this research being conducted in 2016, as now is 2023, more research has been conducting regarding this field of interest. Thus, based on the recent research, I concluded there are two potential things I can add to this research:

1. Adding a feature that makes the prediction more tailored towards differentiating SLI, ASD, and TD.
2. Adding an ML algorithm, specifically Random Forrest, as according to recent research, Random Forrest was proven to be the most accurate (Dewi & Imah, 2020; Qureshi et al., 2023) and Lee did not include Random Forrest but just a decision tree classifier.

Dewi, E. S., & Imah, E. M. (2020). Comparison of machine learning algorithms for Autism Spectrum Disorder Classification. *Proceedings of the International Joint Conference on Science and Engineering (IJCSE 2020)*. https://doi.org/10.2991/aer.k.201124.028

Lee, Z. K. J. (2016). *Use of Natural Language Processing and Machine Learning Techniques to Clinically Assess Children Autism Spectrum Disorder and Specific Language Impairment*.

Qureshi, M. S., Qureshi, M. B., Asghar, J., Alam, F., & Aljarbouh, A. (2023). Prediction and analysis of Autism Spectrum Disorder Using Machine Learning Techniques. *Journal of Healthcare Engineering*, *2023*, 1–10. https://doi.org/10.1155/2023/4853800

Operational Measures

Natural Language Processing (NLP) now aids in identifying key factors indicative of ASD and SLI without the need for manual classification. However, differentiating between ASD and SLI remains challenging due to overlapping and distinct features influencing how a child is categorized. To clarify, SLI is defined as a 'delayed or disordered language pattern in the absence of other sensory, neurological, or intellectual abnormalities,' which can also adversely impact a child's social development and academic achievement. Certain indicative factors like unique characteristics of morphology, syntax, and phonology are similar in both conditions. Yet, ASD typically includes pragmatic impairments and broader developmental difficulties beyond language issues. Differences are also shown here in the research conducted by Schaeffer:

Group	TD	ASD	SLI	Result
Non-word and Sentence Repetition Task	3	1	1	ASD & SLI similar
Finding a simple shape hidden within a more complex shape	2.5	2.5	1	NOT similar
Judge the quantity of two similar objects	2.5	1	3	NOT similar
Subject and verb agreement	2.5	1	3	NOT similar
Find the correct def. Articles	3	1	1	ASD & SLI similar

*Level of Worse: 1 - most to 3 - least

Therefore, the primary goal of this paper is to accurately differentiate between a Typically Developing (TD) child, a child with SLI, and a child with ASD using NLP and Machine Learning (ML) (Lee, 2016, pg. 5). Thus, using the text-based data under the classification of control of TD with 2 experimental data of SLI and ASD from the open source data bank called TalkBank, the study can be conducted.

Hypothesis Development:

The ML algorithm is employed to predict classification labels like TD (Typically Developing), ASD (Autism Spectrum Disorder), or SLI (Specific Language Impairment) based on the input data. Each algorithm employs a distinct method to reach its conclusions, which allows for observation and analysis of how features are classified and detected. Consequently, different ML algorithms yield varying results. This variation implies that each algorithm has its unique strengths. Poliker's strategy capitalizes on this by using multiple learning algorithms simultaneously, aiming to achieve higher predictive capability than what would be possible with each algorithm individually (Lee, 2016, pg. 7).

With X and Y, I am curious about the potential of the Random Forest algorithm in addressing the conclusion from Lee's 2016 study, which stated: “Comparing the confusion matrices of all classifiers, it is observed that the ASD class has the lowest precision and F-measure across all classes, even after feature extraction” (Lee, 2016, pg. 33). I would like to know if precision for ASD can be improved as This interest is accoridng to Dewi and Imah's 2020 research, it’s found that 'based on the specificity and sensitivity value, the Random Forest algorithm with full features is the best algorithm compared to others in classifying ASD in children and adolescents' (Dewi & Imah, 2020, pg. 152). Therefore, I am eager to explore whether Random Forest can not only solve the issues identified in Lee's study but also outperform the SVM, which was the most consistent and best-performing machine learning algorithm in Lee's research, with an accuracy score of 87% +/- 3% (Lee, 2016, pg. 32)."

The Machine Learning Workflow

Model Development

The idea of NLP is that we are allowing the computer to break down the sentence by tokenization – the process of converting a sequence of text into smaller parts, known as tokens – and part-of-speech tagging (POS) to analyze the semantics and associate with a feature to distinguish between TD, ASD, or SLI children. Such a method would allow NLP to be executed to measure such as the mean length of utterance, total number of words, and even sentence complexity for intelligence and more. **As a common trait, many features are usually employed to detect LI in children, and feature with algorithm enhances the precision* (Lee, 2016, pg. 6).

With supervised classification, “detecting patterns is central to NLP as these patterns usually hold new meaning that can be derived from a sentence” (Lee, 2016, pg. 6). As for this research, Lee was using supervised

classification, which means after NLP finds the features, the training was done under supervision to teach what is true to generate models and the corresponding feature sets to complete the prediction phase using the MLs.

1: Feature Selection	2: Training	3: Prediction
Choosing which feature will be used to find patterns (NLP)	Supervised Classification: The model is trained using pairs of labels and feature sets.	the model uses these learned patterns to predict labels for new inputs
	Training based on pairs of labels and feature sets + corpora	
	In the context of a supervised classifier, "corpora" refers to a collection of data used for training the classifier. This data set includes both the input data and their corresponding correct labels.	

Result Presentation & Model Evaluation

With ML, ML algorithm is employed to predict classification labels like TD (Typically Developing), ASD (Autism Spectrum Disorder), or SLI (Specific Language Impairment) based on the input data. Each algorithm employs a distinct method to reach its conclusions, which allows for observation and analysis of how features are classified and detected. Consequently, different ML algorithms yield varying results. This variation implies that each algorithm has its unique strengths. Poliker's strategy capitalizes on this by using multiple learning algorithms simultaneously, aiming to achieve higher predictive capability than what would be possible with each algorithm individually (Lee, 2016, pg. 7). Thus, as my research is based on Lee's research, I would do the similar. Thus, even for the data visualization, with using confusion matrix, I will compare the results of each ML algorithms before and after the feature extraction to each other as well as to the baseline performance as the evaluation(Lee, 2016, pg. 20).

Specifically, with evaluation, precision recall f-measure and recall measures will be measured using this formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Additionally, with specifically additionally adding one more ML algorithm of Random Forest, the parameter settings based on the parameter Random Forest provides will be evaluated for instance like max depth.