Unraveling a Linguistic Enigma: Computational Approach to Understanding Specific Language Impairment in Children

Barto Radman (2062836), Christophe Friezas Gonçalves (2059012), Mario Vella (2085712)

November 2023

1 Introduction

Throughout the early years, namely from the time of their birth to the period of their young adult life, the acquisition of language skills is a critical developmental milestone in a child's life. The development of language skills occurs in the majority of children during the first three years of their lives, despite the differences in the languages being learned, the learning environment and intelligence (Eigsti et al., 2011). However, with severe enough differences, some children display evident difficulties in their language learning abilities. For instance, children diagnosed with autism spectrum disorders (ASD) or hearing loss are denoted by language and communication impairment (Eigsti et al., 2011; Yoshinaga-Itano et al., 1998). These disorders affect the language acquisition abilities of a child, however, their causes are distinct in nature. The onset of ASD is understood through the neurobiological causes within the brain while hearing loss is physically observed by paediatricians so it can be diagnosed as early as possible (Eigsti et al., 2011; Yoshinaga-Itano et al., 1998). The present study will focus on specific speech impairment (SLI), which in contrast is not distinguished by any evident impairments. Children diagnosed with SLI display no developmental issues, normal hearing sensitivity, and normal nonverbal intelligence (Montgomery, 2002). Regardless of seemingly normal function, comprehending and producing new words, using complex sentences and disorganized storytelling and grammatical errors (Leonard, 2017) are typical difficulties observed in children with SLI. Numerous possible explanations for the cause of SLI are proposed in the literature (Gray, 2004; Gray, 2005). An example can be found in research conducted by Evans et al. (2009) where language learning is described as a statistical learning mechanism, where language consists of a mental lexicon and mental grammar, interacting with the memory of a child. It is hypothesized that children diagnosed with SLI have impairments in these language learning systems (Evans et al., 2009). The present paper aims to cognitively model known features of children diagnosed with SLI. Various machine learning (statistical) models are trained on data consisting of children's utterances during a storytelling task. The aim of the model is to correctly classify the diagnosis of the child as either normal or SLI. The model will take into account the features which describe the aforementioned difficulties through examining the mean sentence sizes, amount of pauses, interventions by the investigator and so on. It is predicted that the statistical models would be able to correctly distinguish the linguistic profile of children with SLI, providing a more robust understanding of important speech patterns in children with SLI. The literature on SLI also considers the age of children, as SLI is often diagnosed through comparison of a normal developing child (Evans et al., 2009). The paper thus aims to answer a research question: *Do children diagnosed with SLI have distinct language features which distinguish them from normally developing children?* The age of the children are also taken into account as a feature to distinguish potential age related differences.

2 Dataset

The proceeding models will be trained on features extracted from the CHILDES Clinical English Gillam corpus (Gillam et al., 2004). The dataset consists of transcripts of interviews with 173 children with SLI and 498 control children with normal linguistic development. The age of the children ranged from 5 to 11 year old. The children were tasked on creating different stories, based on three predefined topics: McDonalds, aliens and being late for school. In this study 173 SLI entries and 173 randomly chosen control entries will be used to balance the results and train the model.

2.1 Features

As mentioned in the introduction, seven different features were extracted from the dataset. The features include amount of utterances and mean utterance size over the whole interview, length of longest utterance, pauses, interventions by the interviewer, vocabulary size and age. The interviews were cleaned by removing unnecessary annotations (whispering tags, grammar tags, ...), followed by a line by line parse to extract the pause tags as well as count the sentence specific features. The intervention feature takes into account all interaction the interview takes with the child throughout the tasked story delivery. The majority of features focus on the proficiency of sentence creation, given the wide range of deficits created by SLI (Kohnert et al., 2009). Furthermore, Gray (2004) showed that SLI children have particular difficulties in producing language, especially on the word level. The longest utterance length, amount of utterances and mean utterance size, as well as vocabulary size, denote the features tackling this aspect in our models. The pause feature in addition to the intervention feature tackle the complex sentence structure and disorganized storytelling deficit of SLI children (Leonard, 2017). The SLI diagnosis is based on comparison between children of the same age group bringing forth the last feature, age, to identify if said feature improves the discernibility for our models, given its main role in real world diagnosis.

3 Model

In this project, five different machine learning algorithms were used with their combinations of hyperparameters to classify whether a specific test data point was a (Normal) child or a child with Specific Language Impairment (SLI). The models implemented for this task were the Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest, ADABoost, and Logistic Regression Models. In Table 1, the models together with their respective hyperparameters and test accuracies are shown.

Model	**Hyperparameters**	**Test Accuracy**
Logistic Regression	$random\_state = 1$, $max\_iter = 1000$	**0.7857**
Multi-Layer Perceptron	$random\_state = 1$, $max\_iter = 1000$	0.7571
Support Vector Machine	$kernel = rbf$	0.7143
Random Forest	$max\_depth = 4$, $random\_state = 42$	0.6714
ADABoost	$n\_estimators = 40$, $random\_state = 42$	**0.6571**

Table 1: Test Accuracies of Implemented Models

As can be seen in Table 1, the Logistic Regression model performed the best with an accuracy of 0.7857, and the ADABoost performed the worst with an accuracy of 0.6571. As can be seen, the range between the most and least performing models is not big (Range of $0.7857 - 0.6571 = 0.1286$). As mentioned in section 2, the datasets for the Normal and SLI children are of equal size, which makes the models work better. In Figure 1, the Confusion Matrix of the best performing model i.e. Logistic Regression can be seen.

3.1 Feature Importance

Feature importance is the metric displaying how important certain input features are to the required output. Analyzing the importance of such features is critical to answering the main research question, namely which features can be used to successfully discern children with SLI as compared to children with normal language development. In this section, we will be showing the feature importance shown by the ADABoost and the Random Forest Models using their given functions. These can be seen in the Bar Graphs below (Figures 2 and 3). In Figure 2, the feature importance of the ADABoost model can be seen. Through this graph, it can be seen that the Vocabulary Size of the child in question was the most important feature when classifying between Normal and SLI children. The second most important feature is the number of pauses the child took followed by the age of the child. The interpretation of this graph is that although age seems to be one of the top three features, the age of the children would have been used together with another feature to group the different classes of children (Normal

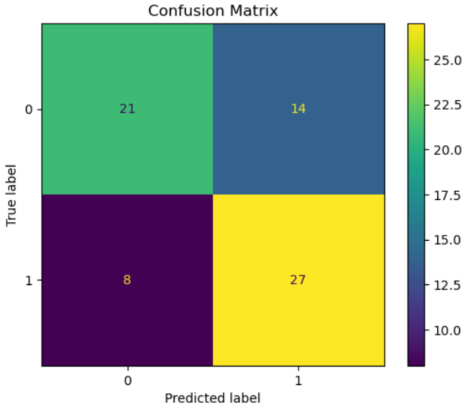

Figure 1: Confusion Matrix: Logistic Regression

and SLI). As mentioned, age of the children is usually taken into account through the comparison of children of similar age with (and without) the disorder. The models might not correctly take into account the comparative use of age as a feature in the diagnosis of speech disorders.

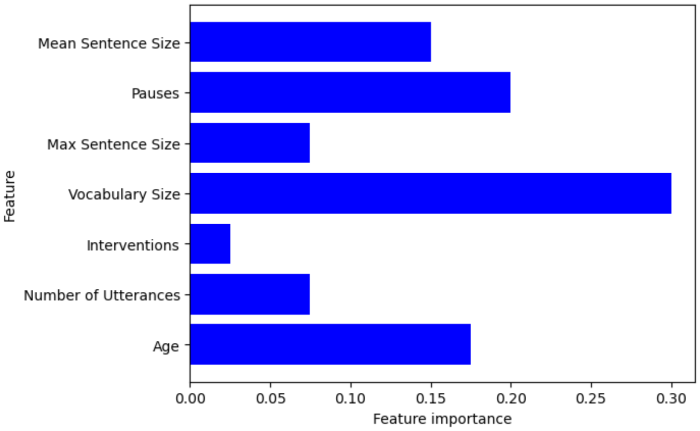

Figure 2: Feature Importance: ADABoost

Furthermore, in Figure 3, one can see the same type of graph for the Random Forest model. In this graph, the main three features responsible for the classification are different barring one. These are the mean sentence size of the speech, the vocabulary size of the child, and the number of utterances the child makes. Comparing these results to the ADABoost model, the mean sentence size and the number of utterances provide a different effect on the classification. These differences could be due to the differences in how these models work. An example of this is that an ADABoost Model combines different weak models that focus on the mistakes made by the previous models and gives a bigger weight to the samples that are incorrectly classified. On the other hand, a Random Forest model constructs multiple decision trees and trains them using different random subsets of the dataset.
It is also important to mention that all of the suspected features, extracted from the dataset, have an sizable influence on the ability to correctly diagnose the child. The results of the model analysis further reinforce the complex nature of SLI, where multitude of aspects regarding the production of speech influence the disorder. The analysis displays evidence for the plausibility of computational models ability to be used in diagnosis in subtle disorders.

4 Conclusion

The goal of the presented computational model was to correctly diagnose children with or without the specific language impairment. Specific language impairment remains an elusive disorder due to the less evident causes and normal development, as compared to other known speech disorders. Thus, taking a computational modeling approach to the diagnosis could help shine light on the importance

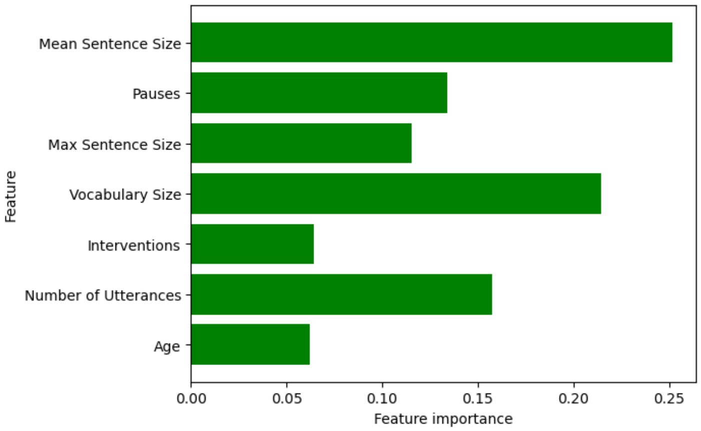

Figure 3: Feature Importance: Random Forest

of salient speech features. The study trained five machine learning models on a database containing utterances of children with SLI and normal development. Using Logistic Regression, the model is able to correctly diagnose the child with 79% accuracy. Interpreting the importance of the specific features contained in the linguistic profile of children with SLI gives insight on possible difficulties evident from production and understanding of speech. Vocabulary size, unsurprisingly, remains as one of the most important indicators of a speech impairment. While the finding is not surprising, as it is expected that children with SLI have a considerably smaller and less variant vocabulary, it nonetheless provides a starting ground for further investigation. Delving deeper into potential causes for the differences in vocabulary size, along with other prominent features, would convey a deeper understanding of SLI. Lastly, it is critical to understand the limitations of machine learning models and what the findings are able to tell us about SLI. First limitation is the data and what is able to be learned from it. While the dataset contains real utterances from children with SLI and normal language development, it is not known whether the model is correctly classifying SLI or just a difference in language in general. It should further be investigated whether models could discriminate SLI among other known speech disorders. The model would likely run into issues as the other disorder could be characterized by similar features, grouping them all into a ”disorder” class. Thus, what we can interpreted from the model is that there is indeed a specific linguistic profile for speech disorders. The models are also limited in discerning the actual causes. The findings reached from computational modeling are not able describe exactly what causes the differences in vocabulary size, or comprehension issues or even the number of pauses. Despite the limitations, computational modeling can be used to foster further research. Correctly classifying between a multitude of different disorders would be a natural next step, along with developing knowledge on potential causes for disorders with less salient impairments. Ultimately, being able to successfully recognize and provide necessary interventions to children with SLI (and other disorders) could assist children in normal linguistic development.

5 References

[1] Eigsti, I.-M., de Marchena, A. B., Schuh, J. M., amp; Kelley, E. (2011). Language acquisition in Autism Spectrum Disorders: A developmental review. Research in Autism Spectrum Disorders, 5(2), 681–691. https://doi.org/10.1016/j.rasd.2010.09.001

[2] Evans, J. L., Saffran, J. R., amp; Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. Journal of Speech, Language, and Hearing Research, 52(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189)

[3] Gillam, R. B., Pearson, N. (2004). Test of Narrative Language. Austin, TX: Pro-Ed Inc.

[4] Gray, S. (2004). Word learning by preschoolers with specific language impairment. Journal of Speech, Language, and Hearing Research, 47(5), 1117–1132. https://doi.org/10.1044/1092-4388(2004/083)

[5] Kohnert, K., Windsor, J., amp; Ebert, K. D. (2009). Primary or “specific” language impairment and children learning a second language. Brain and Language, 109(2–3), 101–111. https://doi.org/10.1016/j.bandl.2008.01.009

[6] Leonard, L. B. (2017). Specific language impairment. Oxford Research Encyclopedia of Psychology. https://doi.org/10.1093/acrefore/9780190236557.013.64

[7] Gillam, R. B. & Pearson, N. (2004). Test of Narrative Language. Austin, TX: Pro-Ed Inc.

[8] Gray, S. (2005). Word learning by preschoolers with specific language impairment. Journal of Speech, Language, and Hearing Research, 48(6), 1452–1467. https://doi.org/10.1044/1092-4388(2005/101)

[9] Montgomery, J. W. (2002). Understanding the language difficulties of children with specific language impairments. American Journal of Speech-Language Pathology, 11(1), 77–91. https://doi.org/10.1044/1058-0360(2002/009)

[10] Yoshinaga-Itano, C., Sedey, A. L., Coulter, D. K., amp; Mehl, A. L. (1998). Language of early- and later-identified children with hearing losPaediatricsics, 102(5), 1161–1171. https://doi.org/10.1542/peds.102.5.1161