2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

# Predicting Severity in People with Aphasia: A Natural Language Processing and Machine Learning Approach

Marjory Day[1], Rupam Kumar Dey[1], Matthew Baucum[1], Eun Jin Paek[2], Hyejin Park[3], Anahita Khojandi[1]

*Abstract*—**Speech language pathologists need an accurate assessment of the severity of people with aphasia (PWA) to design and provide the best course of therapy. Currently, severity is evaluated manually by an increasingly scarce pool of experienced and well-trained clinicians, taking considerable time resources. By analyzing the transcripts from three discourse elicitation methods, this study combines natural language processing (NLP) and machine learning (ML) to predict the severity of PWA, both by score and severity level. By engineering language features from PWA tasks, an unstructured k-means clustering presents distinct aphasia types, showing validity of the selected features. We develop regression models to predict severity scores along with a classification of severity by level (Mild, Moderate, Severe, and Very Severe) to assist clinicians to easily plan and monitor the course of treatment. Our best ML regression model uses a deep neural network and results in a mean absolute error (MAE) of 0.0671 and root mean squared error (RMSE) of 0.0922. Our best classification model uses a random forest and result in an overall accuracy of 73%, with the highest accuracy of 87.5% for mild severity. Our results suggest that using NLP and ML provides an accurate and cost-effective approach to evaluate the severity levels in PWA to consequently help clinicians determine rehabilitation procedures.**

*Keywords*— **Aphasia, Discourse Elicitation Method, Natural Language Processing, Machine Learning, Speech Language Pathology**

## I. Introduction

### A. Background

Aphasia is a language disorder which leads to failure to understand or formulate language because of damage to specific brain regions [1]. Aphasia is commonly caused by a cerebral vascular accident (also known as stroke), specifically in older adults. Other causes of aphasia include brain tumors, brain infections, or neurodegenerative diseases [2] .

Aphasia affects about two million people in the US and 250,000 people in Great Britain [3]. Nearly 180,000 people acquire the disorder every year in the US alone [4]. Any person of any age can develop aphasia, given that it is also caused by a traumatic injury. However, aphasia is more prevalent in middle-aged and older adults. For example, approximately 75% of all strokes occur in individuals over the age of 65. Twenty five percent to 40% of stroke survivors develop aphasia because of damage to the language processing regions of the brain [5].

Assessment of aphasia severity in people with aphasia (PWA) is a vital part to determine which treatment procedure will serve best for an individual with aphasia [6]. Although speech-language pathologists all over the world are trying their best to support aphasia patients, getting a reliable and realistic aphasia severity score (and accordingly aphasia classification based on severity scores) is often a time-consuming and cumbersome task. Successful retrieval of necessary linguistic skills from PWA is a vital part for achieving precise severity scores (or in other words severity classifications).

To assess linguistic skills among PWAs, numerous assessment processes are available. Two of the most prominent processes include Western Aphasia Battery – Revised (WAB-R) and Discourse Analysis (DA) [7]. Although scoring procedures are quick for WAB-R compared to DA, the initial test time may vary roughly from 45 minutes to several hours based on the severity of the PWA which is stressful for people who have recently experienced a stroke and are in the recovery phase [8]. Also, language samples collected through WAB-R are not very reflective of natural interactions. On the other hand, DA collects linguistic skills more successfully both in terms of shorter period (about 15-25 mins) of time and in terms of relevancy to everyday conversation compared with WAB-R. However, it takes much longer time for DA in manual scoring and analysis.

DA assessment generally takes multiple tasks into account from multiple different discourse elicitation methods. The most commonly used tasks are 'Cinderella Story' (narrative), 'Cat Rescue' (expositional) and 'Peanut Butter and Jelly Sandwich' (procedural). Cinderella Story is a story narrative task. The participants are first allowed to look through a picture book of Cinderella before being tasked with recalling as much of the story from the narrative as possible without looking back to the book. Cat Rescue falls into the category of expositional discourse where the participant looks at a picture and is asked to tell a story from it having the visual stimuli in front of them all the time. In procedural discourse, the participant is asked about 'how to make peanut butter and jelly sandwich' with no visual stimuli.

Artificial intelligence (AI) can assist clinicians to make better clinical decisions (e.g., predicting disease severity scores, severity type etc.). Two of the popular AI categories are machine learning (ML) and natural language processing (NLP). ML is the study of computer algorithms that matures automatically through experience using historical data [9] ML algorithms build a model, based on sample data (also known

[1] Marjory Day, Rupam Kumar Dey, Matthew Baucaum, and Anahita Khojandi are with the Department of Industrial and Systems Engineering, University of Tennessee-Knoxville. (Corresponding authors: khojandi@utk.edu, rdey1@vols.utk.edu)

[2] Eun Jin Paek is with the Department of Audiology and Speech Pathology, University of Tennessee-Knoxville.

[3] Hyejin Park is with the Department of Communication Sciences and Disorders, The University of Mississippi.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

as 'training data'). It consequently makes predictions without being explicitly programmed to do so [10]. NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interaction between computers and human language, specifically how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of understanding the contents and contexts of the documents. As ML methods are particularly suited for structured data, NLP can come into play when the data is unstructured.

### B. *Literature Review*

PWA often find it challenging to express themselves to people around them. As a result, their family relationship, social life, and work life are immensely affected. Jothi et al. [11] proposed a speech intelligence system bridging the gap between the PWAs and the society by analyzing the unstructured words or unfinished sentences spoken and predict those to meaningful words/ sentences by considering a dataset of unfinished words. Järvelin & Juhola [12] classified aphasic and non-aphasic people considering three aphasia related classification problems with eight different ML classifiers, with no single classifier meaningfully outperforming the others.

Qin et al. [13] took an end-to-end approach to formulate a binary classification task to differentiate PWAs with high scores (Aphasia Quotient ≥ 90) from those with low scores (Aphasia Quotient < 90). Le et al. [14] proposed an idea of developing an intelligent system capable of providing automatic feedback to the patients about their verbal output during practice session regarding sentence building and picture description.

As the diagnosis and evaluation of aphasia takes significant amounts of resources (e.g., time and effort from both clinicians and patients), Dalton & Richardson [15] performed Main Concept Analysis (MCA) to provide descriptive and comparative statistical information for clinicians and researchers about the performance of a large sample of people not brain injured and PWAs on AphasiaBank discourse tasks. Johnson et al. [16] examined correlation between the use of nouns and verbs in standardized confrontation naming tasks and discourse tasks and found strong correlation between nouns and tasks.

### C. *Objective of the Study and Contributions*

The objective of our study is to investigate whether ML can identify the level of severity of PWAs. To do so, we consider transcripts from three discourse tasks, namely Cinderella Story, Cat Rescue, and Peanut Butter and Jelly Sandwich. Using NLP, we engineer language features to capture the diversity and verboseness of patient discourse. We use unsupervised ML to cluster patients based on these linguistic features; in doing so, we validate our language feature set as conveying meaningful information about patients with aphasia. We then develop ML models to perform regression and classification to predict patients' severity score and levels.

By engineering language features for each discourse task, we provide clinicians with psychometric properties and linguistic characteristics (e.g., core vocabulary and frequency of words). Our main contribution is developing an end-to-end ML pipeline that allows for automated evaluation of PWA to decrease the amount of time and resources needed to analyze aphasia severity, or more broadly, aphasiac speech patterns.

## II. Methods

### A. *Data*

The data used for this project was provided by TalkBank, a database of patient interviews from various Aphasia studies across the country [17]. The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board. The three selected tasks encompass three types of discourse tasks gathered in the aphasia interviews.

As aphasia is recognized after a neural event, we studied only PWA. Our final dataset included the 238 participants with aphasia.

We took the data from its transcript and collected the participant responses into a single dataset. From the initial dataset, we removed gestures and commonly spoken words referred to as "stop words" before moving to feature engineering for our severity ranking models.

### B. *Features*

After removing all stop words, we created a *document-term matrix* from participants' transcripts that tabulated the frequency of each word spoken in each of the three tasks. Using the document-term matrix, we identified the top ten words for each of the three tasks, which are:

**Cinderella Story:** ['cinderella', 'ball', 'go', 'prince', 'two', 'one', 'slipper', 'get', 'dress', 'she's']
**Cat Rescue:** ['cat', 'tree', 'get', 'ladder', 'dog', 'girl', 'man', 'little', 'fire', 'got']
**Peanut Butter and Jelly Sandwich**: ['butter', 'bread', 'peanut', 'jelly', 'put', 'get', 'two', 'take', 'would', 'one']

With the top spoken words above, we found the total occurrences of each top word and the total number of top words spoken for each patient. Based on a previous study of Cinderella Story, PWAs experience a decrease in diversity of words used [18]. Hence, we captured language diversity by engineering features that tracked the following:

- Transcript length, which captures the total number of words spoken by a participant excluding stop words
- Words present from the top 10 words, which captures the number of top 10 words that are spoken by a participant at least once. Participant may use all or none of these top 10 words, resulting in values that range from 0-10
- Count of top words spoken, which captures the number of times each of the present top 10 words are spoken in a transcript.

### C. *Models*

To investigate the validity of the engineered features, we used k-means clustering to identify patient subtypes with unique linguistic patterns. We did not include demographics in the clusters (i.e., aphasia type, severity score determined by WAB-R, age, and gender were excluded), using just transcript length, task top 10, and the count of top words spoken.

Next, we used regression models to predict patients' aphasia severity scores, and classification models to predict each patient's level of impairment. Specifically, to predict patients' aphasia severity scores from their task transcripts we used neural network and random forest regression, both of which are commonly used ML techniques. We then identified the best-performing ML model and used it to classify patients' aphasia severity categories, where combining the severe and very severe categories into a single severe/very severe label, i.e., Mild, Moderate, Severe/Very Severe.

Our base model was a linear regression for all language and demographic features. We developed a neural network model with two hidden layers of 39 units each to account for the total features analyzed. We used the same architecture for both classification and regression tasks, except for the changes in the output layer from one to three nodes to account for the three severity categories. To optimize the models, we used the Adam optimizer with rectified linear unit activation function (ReLU) activation function for the two hidden layers.

For the random forest regressor, we first used all demographic and language features before building a model of the 10 most important features. Each random forest used 1000 trees and maximum depth of 20. The random forest classifier mimicked the random forest regressor with all features.

For clustering models, we used the "elbow" point to determine the final number of clusters. For regression metrics, we considered how the model performed both as mean absolute error (MAE) and root mean squared error (RMSE). The classification used F-score to account for the skewed distribution of categories along with accuracy, precision, and recall for each severity category. All models are implemented in Python programming language [19], specifically using the Pandas [20], scikit-learn [21], Keras [22], NumPy [23], Python regular expression operations [24], and string [25] modules.

## III. RESULTS

Table I provides the demographic information for our PWA. As seen in the table, the data were highly skewed with the majority of aphasic patients falling into the Mild category, or a severity score of greater than 76.

TABLE I: DESCRIPTIVE STATISTICS OF PATIENT DEMOGRAPHICS. COUNT OR AVERAGE WITH STANDARD DEVIATION IS PROVIDED.

| **Feature** | **Data Summary** |
|---|---|
| **N** | 238 |
| **Age** | 61.84 (11.58) |
| **Gender** | 92 females<br>146 males |
| **Aphasia Type:** | |
| **Anomic:** | 110 |
| **Broca** | 56 |
| **Conduction** | 41 |
| **Transcortical Motor** | 9 |
| **Transcortical Sensory** | 1 |
| **Wernicke** | 21 |
| **Severity Category:** | |
| **Mild (>76)** | 123 |
| **Moderate (51-75)** | 95 |
| **Severe/Very Severe (0-50)** | 20 |

Table II provides the list of the top 10 most frequently spoken words for each task, per feature extraction enabled by NLP techniques. As seen in the table, these automatically extracted words generally match the expectation for all three tasks.

TABLE II: LANGUAGE FEATURES BY DISCOURSE TASK.

| ***Word Rank*** | **Top 10 Words by Task** | | |
|---|---|---|---|
| | ***Cinderella Story*** | ***Cat Rescue*** | ***PB&J Sandwich*** |
| 1 | cinderella | cat | butter |
| 2 | ball | tree | bread |
| 3 | go | get | peanut |
| 4 | prince | ladder | jelly |
| 5 | two | dog | put |
| 6 | one | girl | get |
| 7 | slipper | man | two |
| 8 | get | little | take |
| 9 | dress | fire | would |
| 10 | she's | got | one |

Per the elbow method, we used $n = 3$ as the final number of clusters. Table III presents the three clusters per demographics, transcript length, and number of top 10 words spoken.

TABLE III: K-MEANS CLUSTERING BY DEMOGRAPHICS AND AGGREGATED FEATURES FOR EACH TASK.

| **Features** | **K-means Clustering** | | |
|---|---|---|---|
| | ***Cluster 0*** | ***Cluster 1*** | ***Cluster 2*** |
| N in Cluster | 139 | 27 | 72 |
| Age | 61.4 | 61.2 | 63 |
| Severity: | 69.2 | 79.2 | 76 |
| Mild | 59 | 20 | 45 |
| Moderate | 67 | 7 | 20 |
| Severe/Very Severe | 13 | 0 | 7 |
| Female | 53 | 9 | 30 |
| Male | 86 | 18 | 42 |
| Anomic | 51 | 16 | 43 |
| Broca | 53 | 0 | 3 |
| Conduction | 18 | 8 | 15 |
| Transcortical Motor | 9 | 0 | 0 |
| Transcortical Sensory | 1 | 0 | 0 |
| Wernicke | 7 | 3 | 11 |
| Cinderella Length | 513.1 | 3169.6 | 1413.8 |
| Cat Length | 254 | 897.6 | 478 |
| Sandwich Length | 148.3 | 565.8 | 290.4 |
| Cinderella Top 10 | 3.6 | 7.5 | 6.5 |
| Cat Top 10 | 4.8 | 7.7 | 6.2 |
| Sandwich Top 10 | 4.2 | 6.9 | 5.6 |

The three clusters appear to be separated by transcript length, with the most verbose patients in Cluster 1 and the least verbose in Cluster 0. The clustering separated several aphasia types naturally into specific clusters, for example, patients with transcortical motor and Broca aphasia are mostly in Cluster 0.

We have used both neural networks and random forest to predict severity scores for PWA. Table IV presents the regression model performances.

TABLE IV: SEVERITY REGRESSION MODEL RESULTS.

| **Errors** | **Linear Regression** | **Neural Network** | **Random Forest** | |
|---|---|---|---|---|
| | | | ***All Features*** | ***Top 10 Features*** |
| MAE | 0.095 | **0.067** | 0.088 | 0.090 |
| RMSE | 0.122 | **0.092** | 0.108 | 0.111 |

While specific raw scores are important, the severity category is also vital to rehabilitation treatment courses. For classification, we have classified the whole set of data into three broad categories i.e., Mild, Moderate, and Severe/ Very Severe. The random forest classifier outperformed the neural network. Table V presents the classification results.

TABLE V: SEVERITY CATEGORY CLASSIFICATION MODEL RESULTS. [ACC.: ACCURACY]

| **Model Metrics** | **Neural Network** | | | **Random Forest** | | |
|---|---|---|---|---|---|---|
| | ***Mild*** | ***Moderate*** | ***Severe /Very Severe*** | ***Mild*** | ***Moderate*** | ***Severe /Very Severe*** |
| Precision | 0.33 | 0.61 | 1 | 0.78 | 0.63 | 1 |
| Recall | 0.5 | 0.54 | 0.17 | 0.88 | 0.67 | 0.33 |
| F-score | 0.40 | 0.57 | 0.29 | 0.82 | 0.65 | 0.5 |
| Acc. | 0.54 | 0.50 | 0.17 | 0.88 | 0.67 | 0.33 |

## IV. DISCUSSION

The k-means clustering results show the validity of the language feature considering the aphasia type separations. By using the language features we engineered from the Cinderella Story, Cat Rescue, and Peanut Butter and Jelly Sandwich discourse tasks, we used random forest and neural networks as regression models to predict severity scores. In addition, we used these models to predict each patient's level of impairment. The MAE of less than 0.07 with a neural network regressor and overall accuracy of greater than 73% for the random forest classification model indicate that ML is a viable solution to decrease the burden on clinicians to analyze transcript data. The best classification and regression models also suggest that different ML approaches should be used depending on the type of prediction required.

The results of this study encourage more exploration into new language features across more tasks. This will grant clinicians more access to transcript analysis without manual grammaticalization or lengthy aggregation. Future studies should look to capture diversity of speech or repetitions through feature selection or detecting levels of severity within a category. In addition, future research needs to investigate detection for severe/very severe patients as the data in this study were mostly skewed toward mild and moderate patients. Through ML, clinicians can not only detect the properties elicited from each discourse task, such as work frequency or core vocabulary, but also can assess changes in severity during a course of treatment more frequently.

## V. CONCLUSION

In this study, we developed ML techniques to decrease the time required by clinicians to determine severity score and impairment level in PWA. We first extracted features using NLP from three common discourse tasks, namely, Cinderella Story, Cat Rescue and Peanut Butter and Jelly Sandwich in PWA. Consequently, through k-means clustering and ML regressions and classification models (mainly neural network and random forest), we showed the validity and efficacy of engineered language features to create meaningful severity scores and levels. These results underscore the potential of NLP and ML to aid with assessments in PWA and support the need for further exploration into using the techniques to streamline and facilitate the assessment process

## REFERENCES

[1] A. R. Damasio, "Aphasia," *New England Journal of Medicine,* vol. 326, no. 8, pp. 531- 539, 1992.

[2] A. S.-L.-H. A. (ASHA), "Aphasia," [Online]. Available: https://www.asha.org/public/speech/disorders/Aphasia/. [Accessed 02 05 2021].

[3] N. A. Association, "Aphasia Statistics," [Online]. Available: https://www.aphasia.org/aphasia-resources/aphasia-statistics/. [Accessed 02 05 2021].

[4] N. A. Association, "Aphasia Fact sheet," 31 July 2018. [Online]. Available: https://www.aphasia.org/aphasia-resources/aphasia-factsheet/. [Accessed 02 05 2021].

[5] N. A. Association, "Aphasia FAQs," 30 October 2019. [Online]. Available: https://www.aphasia.org/aphasia-faqs/. [Accessed 03 05 2021].

[6] L. L. Q. B. H. &. L. S. Zheng, "Speech Emotion Recognition Based on Convolution Neural Network Combined with Random Forest," *Chinese Control and Decision Conference (CCDC),* pp. 4143-4147, June 2018.

[7] D. F. M. H. A. D. S. G. R. J. &. M. B. Fromm, "Discourse Characteristics in Aphasia beyond the Western Aphasia Battery Cutoff.," *American Journal of Speech-Language Pahtology,* vol. 26, no. 3, pp. 762-768, 2017.

[8] L. S. Turkstra, C. Coelho and M. Ylvisaker, "The Use of Standardized Tests for Individuals with Cognitive-Communication Disorders," *Seminars in Speech and Language,* vol. 26, no. 04, pp. 215-222, 2005.

[9] T. Mitchell, Machine Learning, New York: McGraw Hill, 1997.

[10] J. R. Koza, F. H. Bennett, D. Andre and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design '96*, Dordrecht, Springer, 1996, pp. 151-170.

[11] K. R. Jothi, V. L. Mamatha, S. B. B. and P. Yawalkar, "Speech Intelligence Using Machine Learning for Aphasia Individual," in *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, 2019.

[12] A. Järvelin and M. Juhola, "Comparison of machine learning methods for classifying aphasic and non-aphasic speakers," *Computer Methods and Programs in Biomedicine,* vol. 104, no. 3, pp. 349-357, 2011.

[13] Y. Qin, Y. Wu, T. Lee and A. P. H. Kong, "An End-to-End Approach to Automatic Speech Assessment for Cantonese-speaking People with Aphasia," *Journal of Signal Processing Systems,* vol. 92, no. 8, pp. 819-830, 2020.

[14] D. Le, K. Licata, E. Mercado, C. Persad and E. M. Provost, "Automatic analysis of speech quality for aphasia treatment," in *International Conference on Acoustics Speech and Signal Processing*, Florence, 2014.

[15] S. G. Dalton and J. D. Richardson, "A Large-Scale Comparison of Main Concept Production Between Persons With Aphasia and Persons Without Brain Injury," *American Journal of Speech-Language Pathology,* vol. 28, no. 1S, pp. 293-320, 2019.

[16] K. M. Johnson, J. Kurland, J. Parker, D. Fromm and B. McWhinney, "Nouns and Verbs in Naming and Storytelling Tasks in Aphasia: Verbs are Another Story," in *42nd annual Clinical Aphasiology Conference*, 2012.

[17] B. F. D. F. M. &. H. A. MacWhinney, "AphasiaBank: Methods for studying discourse," *Aphasiology, 25,* pp. 1286-1307, 2011.

[18] B. F. D. H. A. F. M. W. H. MacWhinney, "Automated Analysis of the Cinderella Story," *Aphasiology,* vol. 24, pp. 856-868, 2010.

[19] Python Core Team , "Python: A dynamic, open source programming language.," 2015. [Online]. Available: https://www.python.org/.

[20] W. M. j. J. V. d. B. T. A. P. C. …. M. M. Jeff Reback, "pandas-dev/pandas: Pandas 1.0.3 (Version v1.0.3)," 18 March 2020. [Online]. Available: http://doi.org/10.5281/zenodo.3715232.

[21] P. e. al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[22] F. Chollet, "Keras," *GitHub repository,* vol. https://github.com/fchollet/keras, 2015.

[23] C. M. K. v. d. W. S. e. a. Harris, "Array programming with NumPy," *Nature 585,* pp. 357-362, 2020.

[24] G. Van Rossum, The Python Library Reference, release 3.8.2, Python Software Foundation, 2020.

[25] GitHub, "Python/cpython," [Online]. Available: https://github.com/python/cpython/blob/master/Doc/library/string.rst.