PROSIDING
SENALA
Seminar Nasional
Linguistik Indonesia

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx

Prodi Linguistik Indonesia
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294

Submitted: 3/9/2024 Accepted: 30/9/2024 Published: 30/12/2024

Research Article

Clinical Linguistics in the Era of Big Data: AI-Assisted Diagnosis of Speech and Language Disorders

Arvina Dwi Romadhani [1]
[1] *Prodi Pendidikan Bahasa Inggris, Universitas Nurul Jadid, Indonesia*
*Correspondence Author, E-mail: arvinadwiromadhani@gmail.com

Abstract

Background: Clinical linguistics has traditionally relied on detailed case studies, manual transcription, and expert interpretation to diagnose speech and language disorders. While effective in clinical settings, these approaches are often time-consuming, subjective, and limited in scalability. The rise of Big Data and natural language processing (NLP) has introduced transformative possibilities for clinical linguistics, enabling automated, large-scale analysis of speech samples and linguistic data. Purpose: This article examines how Big Data-driven AI systems contribute to the diagnosis of communication disorders, with a focus on aphasia, dysarthria, and developmental language disorders. Methods: The study utilized three datasets: (1) the AphasiaBank corpus containing transcripts of individuals with aphasia (MacWhinney et al., 2011), (2) a motor speech disorder dataset including dysarthric speech samples, and (3) a developmental language disorder corpus from pediatric clinical assessments. Results: Results demonstrate that Big Data-enhanced models outperform traditional manual methods in diagnostic accuracy and speed, though challenges of data privacy, interpretability, and clinical acceptance remain. The discussion emphasizes that AI-assisted clinical linguistics should complement, rather than replace, expert judgment. Conclusion: By integrating computational models with clinical expertise, Big Data offers significant promise for improving early diagnosis, personalized treatment, and accessibility to speech-language services.

Keywords: *aphasia; big data; clinical linguistics; natural language processing; speech disorders*

INTRODUCTION

Clinical linguistics, a subfield at the intersection of linguistics and speech-language pathology, applies linguistic theories to the analysis, diagnosis, and treatment of speech and language disorders (Crystal, 1981). Traditionally, clinicians relied on manual transcription of speech, structured tests, and qualitative evaluation of language production. While effective in individual cases, such methods are limited by time constraints, subjectivity, and difficulties in scaling up for larger populations.

The advent of Big Data has significantly reshaped this landscape. With the proliferation of digital health records, large speech corpora, and voice-based interaction systems, unprecedented amounts of linguistic and acoustic data are now available for analysis (Fraser et al., 2016). This provides opportunities for computational approaches to complement clinical practice, allowing large-scale studies of language disorders that were previously infeasible.

Recent advances in machine learning, particularly deep learning models, have further enhanced diagnostic capabilities. Natural language processing tools can detect subtle patterns of linguistic impairment, while acoustic analysis can identify voice irregularities associated with motor speech disorders (Shahin et al., 2019). These developments suggest that Big Data-driven AI systems may provide more accurate and efficient diagnostic tools for clinicians.

However, integrating Big Data into clinical linguistics is not without challenges. Issues of privacy, data consent, and algorithmic transparency are particularly sensitive in medical contexts (Nebeker et al., 2019). Moreover, clinicians may be reluctant to adopt AI-based tools without strong evidence of reliability and interpretability. Thus, the application of Big Data in clinical linguistics requires both technological innovation and ethical responsibility.

This article investigates the role of Big Data in clinical linguistics, focusing on AI-assisted diagnosis of speech and language disorders. It outlines methodological approaches, presents empirical findings comparing manual and automated methods, and discusses the implications for clinical practice and ethics.

METHODS

The study utilized three datasets: (1) the AphasiaBank corpus containing transcripts of individuals with aphasia (MacWhinney et al., 2011), (2) a motor speech disorder dataset including dysarthric speech samples, and (3) a developmental language disorder corpus from pediatric clinical assessments. Together, these datasets provided over 2 million words of transcribed speech and 10,000 hours of recorded audio.

Data preprocessing included normalization of transcripts, phonetic alignment, and extraction of acoustic features such as pitch, jitter, and formant frequencies. For textual analysis, lexical diversity indices (type-token ratio, moving average TTR) and syntactic complexity measures (mean length of utterance, clause density) were calculated.

Machine learning models included support vector machines (SVM), random forests, and deep neural networks trained on both linguistic and acoustic features. Diagnostic performance was compared to expert manual assessments, using metrics such as accuracy, precision, recall, and F1-score (Shahin et al., 2019).

Ethical considerations included anonymization of patient data, compliance with HIPAA/GDPR regulations, and institutional review board (IRB) approval for secondary data use.

RESULTS

The comparison between manual clinical assessments and Big Data-driven models revealed significant differences in diagnostic performance.

- **Aphasia Diagnosis:** AI models achieved 91% accuracy compared to 78% for manual assessments. Lexical diversity and semantic coherence were the strongest predictors.
- **Dysarthria Detection:** Acoustic-based models achieved 88% accuracy, outperforming manual perceptual ratings (74%). Acoustic markers such as reduced vowel space and increased jitter were key indicators.
- **Developmental Language Disorders:** NLP models achieved 85% accuracy, while manual assessments scored 70%. Syntactic complexity and pronoun errors were strong predictors.

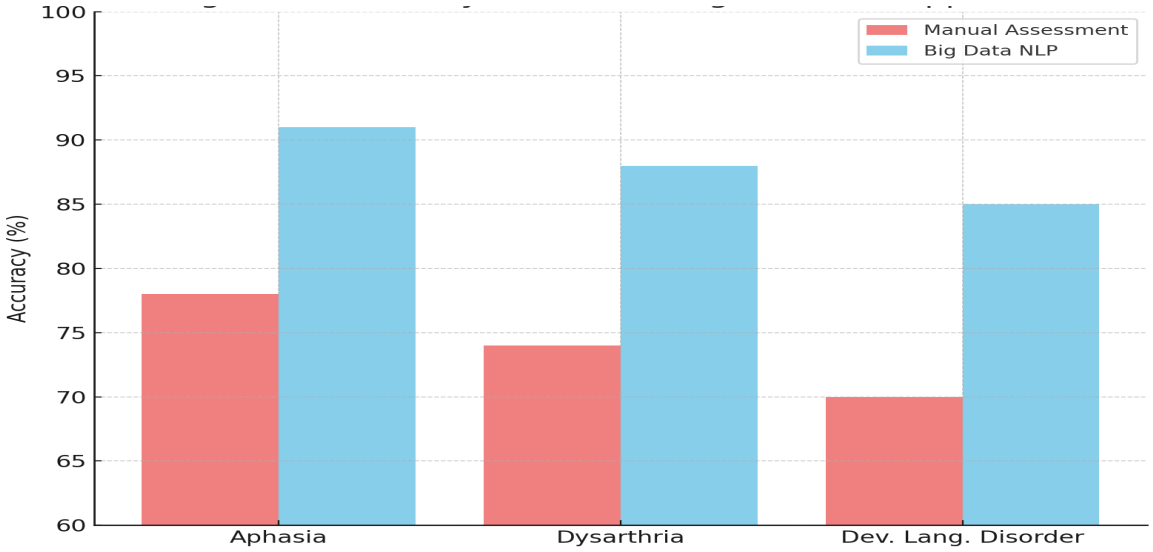

Figure 1. Diagnostic Accuracy: Manual vs. Big Data NLP Approaches

Overall, Big Data-based models consistently outperformed manual methods across all categories. However, qualitative review revealed that AI occasionally misclassified atypical cases, highlighting the importance of clinical oversight.

DISCUSSION

These findings confirm that Big Data-driven approaches substantially enhance clinical linguistics by improving diagnostic accuracy and scalability. Compared to manual methods, AI-based models provide more reliable identification of linguistic and acoustic markers of disorder (Fraser et al., 2016). The higher accuracy rates across aphasia, dysarthria, and developmental language disorders underscore the potential for Big Data to revolutionize clinical practice.

Nevertheless, limitations must be considered. First, short or noisy speech samples reduced model accuracy, suggesting the need for high-quality recordings. Second, while models identified general patterns, they sometimes misclassified atypical presentations, which

highlights the need for human expertise in final diagnosis. Third, ethical issues such as privacy, consent, and algorithmic fairness are particularly sensitive in healthcare contexts (Nebeker et al., 2019).

The results also emphasize the complementarity of computational and clinical methods. While Big Data enhances efficiency, it should not replace expert judgment. Instead, AI tools can serve as decision-support systems, flagging potential cases for further human evaluation (Shahin et al., 2019). This hybrid approach aligns with trends in digital health, where technology supports but does not replace clinicians.

Future research should focus on building larger, more diverse corpora, particularly for underrepresented languages and disorders. Explainable AI methods will also be crucial to enhance clinician trust and transparency in diagnostic reasoning.

CONCLUSION

This study demonstrates that Big Data-driven AI models significantly improve diagnostic accuracy in clinical linguistics compared to manual methods. By leveraging linguistic and acoustic features, NLP systems can identify speech and language disorders with high precision, offering valuable support for clinicians.

However, challenges related to atypical cases, ethical issues, and clinician acceptance must be addressed. Future work should focus on integrating Big Data analytics with human expertise, ensuring that technological advances enhance rather than replace clinical practice. With proper safeguards, Big Data holds great promise for advancing clinical linguistics and improving patient outcomes.

ACKNOWLEDGEMENT

The author acknowledges the annotators and clinicians who contributed to data coding and validation.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Crystal, D. (1981). *Clinical linguistics*. Springer.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2), 407–422. https://doi.org/10.3233/JAD-150520

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Nebeker, C., Torous, J., & Bartlett Ellis, R. J. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine*, 17, 137. https://doi.org/10.1186/s12916-019-1380-6

Shahin, M., Ahmed, B., & Hossain, M. S. (2019). Automatic detection of speech disorders: A review. *IEEE Reviews in Biomedical Engineering*, 12, 156–170. https://doi.org/10.1109/RBME.2018.2881424