International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 4, July – August 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Neural Language Processing for Odia: Challenges, Transformer-Based Approaches, and Future Directions for Low-Resource Language Technologies

Dr. Santosh Kumar Nayak*
*Assistant Professor of Odia, O.E.S.-I
P.G. Department of Odia, Government Autonomous College, Anugola, Odisha-759143, India
Email: skn16031991@gmail.com ORCID ID- https://orcid.org/0000-0003-0089-9098

----------**********----------

Abstract:

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has transformed human-computer interaction across major world languages. However, low-resource languages such as Odia continue to remain underrepresented in computational linguistics due to limited digital resources, inadequate annotated corpora, complex morphology, and insufficient language-specific computational tools. This article critically examines the current status of Odia Language Processing (OLP), emphasizing transformer-based neural architectures, corpus development, machine translation, speech technologies, and language resource creation. It evaluates the linguistic characteristics that distinguish Odia from other Indo-Aryan languages and analyses the computational challenges associated with script processing, morphology, syntax, semantics, and code-mixed communication. Drawing upon recent developments in multilingual language models, the paper proposes an integrated framework for future Odia NLP research combining linguistic knowledge with deep learning methodologies. The study argues that sustainable advancement of Odia language technologies requires interdisciplinary collaboration among linguists, computer scientists, governmental agencies, and digital humanities researchers. The article contributes both theoretically and practically by outlining research priorities capable of strengthening the digital future of the Odia language.

Keywords — **Natural Language Processing; Low-Resource Languages; Transformer Models; Machine Translation; Computational Linguistics**

----------**********----------

I. INTRODUCTION

Natural Language Processing (NLP) has emerged as one of the most influential interdisciplinary domains connecting linguistics, computer science, artificial intelligence, cognitive science, and information technology. Modern NLP systems have revolutionized communication through intelligent search engines, machine translation, speech recognition, question-answering systems, conversational agents, and generative artificial intelligence. Despite remarkable progress for English and other high-resource languages, many regional languages continue to face technological marginalization owing to inadequate computational resources. Odia, one of India's Classical Languages with a literary tradition spanning more than fifteen centuries, represents one such language requiring focused computational attention.

Odia possesses a unique script derived from the Brahmi family, rich inflectional morphology, productive compounding, flexible word order, and extensive lexical borrowing from Sanskrit, Persian, Arabic, Portuguese, and English. These linguistic characteristics present substantial computational challenges for automated language understanding. Although government initiatives such as Digital

India and the National Language Translation Mission have accelerated multilingual computing, Odia language processing still lags behind Hindi, Bengali, Tamil, and Telugu in terms of publicly available corpora, pretrained language models, semantic resources, and evaluation benchmarks.

Recent advances in transformer-based neural networks, multilingual embeddings, and large language models have created unprecedented opportunities for computational development of low-resource languages. These technologies can significantly improve machine translation, sentiment analysis, named entity recognition, automatic summarization, speech technologies, and educational applications for Odia. Consequently, an integrated investigation of contemporary Odia Language Processing becomes both academically significant and socially imperative.

II. REVIEW OF LITERATURE

The foundation of computational linguistics was established through the pioneering works of Chomsky (1957), Jurafsky and Martin (2024), Manning and Schütze (1999), and Bird, Klein, and Loper (2009). Their theoretical frameworks continue to guide modern NLP research involving syntactic parsing, statistical language modeling, and computational semantics. Research on Indian language processing gained momentum through initiatives by the Technology Development for Indian Languages (TDIL), the Indian Language Corpora Initiative (ILCI), and various projects funded by the Ministry of Electronics and Information Technology (MeitY). Studies by Bharati, Chaitanya, and Sangal (1995) demonstrated the applicability of Paninian grammar to Indian language parsing, providing an indigenous computational framework. More recent studies have investigated multilingual transformer architectures such as mBERT, XLM-RoBERTa, IndicBERT, and MuRIL, which substantially improve cross-lingual transfer learning for low-resource Indian languages. However, research specifically dedicated to Odia remains comparatively sparse. Existing investigations primarily address optical character recognition (OCR), Unicode standardization, morphological analysis, and statistical machine translation. Comprehensive research integrating linguistic analysis with contemporary neural architectures remains limited, thereby justifying the present study.

III. OBJECTIVES

The present study aims:

1. To examine the linguistic characteristics influencing Odia Language Processing.
2. To evaluate recent neural approaches applicable to Odia NLP.
3. To identify major computational challenges affecting low-resource language technologies.
4. To propose an integrated research framework for future Odia language processing.

IV. RESEARCH QUESTIONS

- Why does Odia remain a low-resource language despite its classical status?
- Which linguistic properties complicate computational processing?
- How can transformer-based multilingual models improve Odia NLP?
- What future research directions should be prioritized?

V. METHODOLOGY

The study adopts a qualitative descriptive research methodology supplemented by analytical and comparative approaches. Secondary data have been collected from peer-reviewed journals, conference proceedings, books, digital repositories, government reports, and recent NLP research. Contemporary neural language models, multilingual transformer architectures, and Indian language computational frameworks are comparatively analysed to evaluate their applicability to Odia. The study also incorporates linguistic analysis concerning phonology, morphology, syntax,

semantics, and orthography to identify computational bottlenecks.

VI. LINGUISTIC CHARACTERISTICS AFFECTING ODIA LANGUAGE PROCESSING

Unlike many Western languages, Odia demonstrates several linguistic complexities that significantly influence computational processing. The script consists of rounded graphemic structures resulting from historical manuscript traditions. Conjunct consonants, vowel modifiers, ligatures, and orthographic variations increase the difficulty of character segmentation and optical character recognition. Morphologically, Odia exhibits extensive inflection through suffixation. Nouns inflect for number and case, whereas verbs undergo modifications based on tense, aspect, mood, person, and honorificity. Numerous derivational suffixes generate lexical diversity from a single root, producing substantial vocabulary variation. Syntactically, Odia generally follows a Subject–Object–Verb (SOV) order, although pragmatic factors often permit flexible constituent movement. Long-distance dependencies, ellipsis, and implicit subjects complicate dependency parsing and semantic interpretation. Lexically, Odia incorporates Tatsama, Tadbhava, Desaja, and foreign lexical items, producing multiple synonymous expressions. This lexical diversity presents challenges for word embedding generation, semantic similarity estimation, and contextual disambiguation.

VII. NEURAL APPROACHES TO ODIA LANGUAGE PROCESSING

Traditional NLP relied heavily upon rule-based systems and statistical learning. While these methods achieved moderate success, they struggled with complex morphology and contextual ambiguity. Transformer architectures have fundamentally altered NLP by replacing sequential processing with self-attention mechanisms capable of modelling long-range contextual dependencies. Multilingual transformer models pretrained on hundreds of languages enable transfer learning, allowing low-resource languages to benefit from knowledge acquired from larger multilingual corpora. For Odia, transformer-based approaches offer considerable advantages:

- First, contextual embeddings effectively distinguish polysemous words appearing in different semantic environments.
- Second, subword tokenization handles rare and morphologically complex words without requiring exhaustive dictionaries.
- Third, multilingual transfer learning reduces dependence on extremely large Odia-specific datasets.
- Fourth, fine-tuned transformer models substantially improve text classification, sentiment analysis, question answering, and document summarization.

The emergence of Large Language Models further extends these capabilities by integrating reasoning, multilingual generation, dialogue systems, and contextual understanding into unified architectures.

VIII. MAJOR APPLICATIONS OF ODIA LANGUAGE PROCESSING

The practical applications of Odia NLP extend across numerous domains. Machine Translation remains one of the most socially significant applications. High-quality translation between Odia and English or Hindi promotes educational accessibility, governmental communication, and digital inclusion. Speech Recognition systems facilitate voice-controlled interfaces, digital assistants, accessibility technologies, and smart governance for native speakers. Optical Character Recognition enables digitization of historical manuscripts, palm-leaf texts, newspapers, and archival documents, thereby preserving Odisha's literary heritage.

Named Entity Recognition supports information retrieval, digital journalism, legal analytics, and knowledge graph construction.

Sentiment Analysis assists public opinion mining from social media, political communication, customer feedback, and market intelligence. Educational technologies can automatically evaluate student responses, generate learning materials, provide pronunciation assistance, and support language acquisition. Healthcare applications include multilingual medical consultation systems capable of assisting rural populations with limited English proficiency.

IX. CHALLENGES IN ODIA LANGUAGE PROCESSING

Despite recent progress, several obstacles continue to hinder computational advancement.

- The foremost limitation concerns insufficient annotated corpora. Deep learning algorithms require millions of labelled examples, whereas publicly available Odia datasets remain comparatively small.
- Secondly, the absence of comprehensive lexical databases such as WordNet, FrameNet, semantic role corpora, and discourse annotations restricts higher-level language understanding.
- Thirdly, code-mixed communication involving Odia, English, and Hindi has become increasingly common across digital platforms. Mixed scripts and transliterated expressions complicate tokenization and language identification.
- Another significant challenge involves dialectal diversity. Regional variations across coastal, western, southern, and northern Odisha produce phonological and lexical differences requiring dialect-aware computational models.

Furthermore, computational resources, research funding, interdisciplinary collaboration, and open-source language infrastructure remain inadequate compared to globally dominant languages.

X. PROPOSED INTEGRATED FRAMEWORK FOR FUTURE ODIA NLP

Future Odia Language Processing should adopt an integrated research ecosystem comprising five interconnected layers.

- The first layer involves large-scale corpus creation through digitization of literature, newspapers, government documents, folklore, educational resources, and social media.
- The second layer focuses on linguistic annotation incorporating morphology, syntax, semantics, discourse, and pragmatic information prepared by trained linguists.
- The third layer emphasizes transformer-based multilingual pretraining specifically optimized for Odia using balanced datasets.
- The fourth layer integrates multimodal AI combining text, speech, handwriting, and image processing for comprehensive language understanding.

Finally, community-driven open-source platforms should continuously update language resources, evaluation benchmarks, and pretrained models through collaborative participation among universities, government agencies, industry, and independent researchers. Such an ecosystem would substantially reduce the technological disparity between Odia and high-resource languages.

XI. FINDINGS

The study reveals several important findings.

- Odia possesses rich linguistic complexity that simultaneously enriches literary expression and complicates computational modelling.
- Transformer-based multilingual neural architectures significantly outperform traditional statistical approaches for low-resource languages.
- Resource scarcity rather than algorithmic limitation remains the principal barrier to rapid progress in Odia NLP.

- Future development requires balanced integration of linguistic expertise, artificial intelligence, corpus engineering, and governmental policy support.
- Collaborative open-source initiatives are likely to accelerate innovation more effectively than isolated institutional projects.

XII. CONCLUSION

Odia Language Processing stands at a critical historical juncture. While the language possesses an extraordinary literary heritage and constitutional recognition as a Classical Language, its computational infrastructure remains comparatively underdeveloped. Recent advances in transformer architectures, multilingual representation learning, and generative artificial intelligence provide unprecedented opportunities to overcome longstanding technological limitations. The future of Odia NLP depends not merely upon increasingly sophisticated algorithms but upon systematic corpus development, interdisciplinary research, linguistic documentation, open-source collaboration, and sustained institutional investment. Computational linguistics must therefore evolve as a collaborative enterprise where traditional linguistic scholarship and contemporary artificial intelligence complement rather than compete with each other. Such an integrated approach will not only preserve the linguistic richness of Odia but also ensure its active participation in the emerging multilingual digital knowledge society. Advancing Odia language technologies is thus both a scientific challenge and a cultural responsibility, ensuring that one of India's oldest literary traditions flourishes in the age of intelligent machines.

REFERENCES

[1] Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, 1995.

[2] Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[3] Brown, Peter F., et al. "The Mathematics of Statistical Machine Translation." *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 263–311.

[4] Chomsky, Noam. *Syntactic Structures*. Mouton, 1957.

[5] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[6] Doshi, Sahil, et al. "AI4Bharat IndicBERT: A Multilingual Language Model for Indian Languages." *arXiv*, 2021.

[7] Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing*. 3rd ed., draft, 2024.

[8] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International.

[9] Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[10] Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] Wolf, Thomas, et al. "Transformers: State-of-the-Art Natural Language Processing." *Proceedings of EMNLP: System Demonstrations*, 2020, pp. 38–45.

[12] Xue, Linting, et al. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." *Proceedings of NAACL-HLT*, 2021.

