

A COMPREHENSIVE PHILOLOGICAL ANALYSIS TOOL USING AI

Project Reference No.: 49S_BE_2045

College : Bangalore Technological Institute, Bengaluru
Branch : Department of Artificial Intelligence & Machine Learning
Guide : Dr. M A Dorai Ranga Swamy
Student(s) : Ms. Anjali.K.S
Mr. C D Subramanya Iyengar
Mr. Dhanush R

Keywords:

Philological Analysis, Ancient Text Processing, Historical Linguistics, Deep Learning, Neural Networks, Natural Language Processing (NLP), Text Summarization, Semantic Analysis, Syntactic Parsing, Data Preprocessing, Feature Extraction, Convolutional Neural Networks (CNN), Transformer Models, Pre-trained models.

Introduction:

Philology is the interdisciplinary study of language in its historical and literary context that serves as a cornerstone for understanding and preserving cultural heritage. In South Asian studies, it is especially crucial for interpreting classical languages such as Sanskrit and Proto-Dravidian, which carry vast repositories of religious, scientific, and literary knowledge. Traditional philological work, however, is slow and challenging due to fragile manuscripts, script diversity, orthographic variations, and the deteriorated condition of ancient documents. The project proposes an integrated AI system designed specifically for Sanskrit and Proto-Dravidian philological analysis. Semantic embeddings and diachronic models help retain cultural and contextual meaning during translation. The system also supports multilingual script analysis, emotional and metaphor detection, and evaluation of interpretive variations as shown in fig 1.



Figure 1: Variations of the word 'dal' in different languages across regions

Many manuscripts are partially damaged, handwritten in varying styles, or preserved in deteriorated forms, making them difficult to digitize, read, or interpret accurately using manual methods. There is a critical need for an automated, scalable, and reliable AI-based system that can recognize, restore, analyze, and translate ancient manuscripts while preserving their linguistic, semantic, and cultural integrity. This project aims to address these challenges by developing an integrated AI framework capable of performing OCR, linguistic parsing, semantic analysis, and text restoration specifically tailored for Sanskrit and Proto Dravidian philological research.

Objectives:

The primary objective of this work is to develop an AI-powered system capable of:

1. Digitizing
2. Recognizing
3. Restoring
4. Interpreting ancient manuscripts
5. Provide greater accuracy, efficiency, and scalability in analysing the text

Methodology:

The proposed methodology integrates multiple advanced artificial intelligence techniques to enable comprehensive philological analysis. It begins with the systematic collection of diverse textual data, including scanned manuscripts, inscriptions, and historical documents, which often contain complex scripts and degraded content. These inputs are processed using CNN-based models that specialize in script recognition by extracting intricate visual patterns from images. Once the text is recognized, NLP pipelines are applied for linguistic parsing, including tokenization, part-of-speech tagging, and syntactic analysis, enabling a deeper understanding of the textual structure.

Further, semantic extraction techniques are employed to derive contextual meaning from the processed data. For handling multilingual content, Transformer-based architectures are utilized for machine translation, ensuring accurate and context-aware language conversion. To address challenges such as incomplete or damaged texts, sequence-generation networks like LSTMs are used to predict and reconstruct missing segments. Additionally, advanced models such as BART and GPT enhance this restoration process by generating coherent and contextually appropriate text.

The entire system is designed as a modular pipeline consisting of stages such as data collection, preprocessing, model training, validation, and deployment. Each stage operates independently while maintaining seamless integration with others, ensuring flexibility and scalability. Preprocessing focuses on cleaning, normalization, and formatting of data, while the training phase leverages curated datasets to improve model accuracy.

Result and Conclusion:

In conclusion, A Comprehensive Philological Analysis Tool using AI demonstrates how modern computational techniques can significantly enhance traditional linguistic and textual studies. By integrating modules for language detection, text restoration, semantic interpretation, transliteration, and deep linguistic evaluation, the system offers an automated

workflow that reduces manual effort while improving accuracy and consistency. This project shows the potential of AI to handle complex philological tasks that once required extensive expertise and time.

Through machine-learning models, advanced preprocessing, and structured analysis pipelines, the tool can identify corrupted content, restore missing text, interpret meaning, and provide a comprehensive philological report. Each module works cohesively to ensure that even damaged or historically challenging manuscripts can be analyzed with clarity and precision. This capability not only supports researchers but also opens doors for educators, students, and digital humanities practitioners to explore ancient or multilingual texts more effectively.

Overall, the project establishes a robust foundation for future advancements in AI-driven philology. With its modular architecture, user-friendly interface, dataset explorer, and media functionalities, the system is well-positioned for expansion into larger datasets, more languages, and additional scholarly tools. As a whole, it contributes meaningfully to cultural preservation, linguistic research, and the broader field of intelligent text analysis.

References:

1. Aditi Moudgil et al., "CNN-LSTM Based Approach for Recognition of Devanagari Manuscripts", Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI), IEEE, 2022
2. Hamid et al., "Deep Learning-Based Approach for Dating Historical Manuscripts: A Study of CNNs Applied to Paleographic Analysis", Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019
3. A. Sulaiman et al., "A Database for Degraded Arabic Historical Manuscripts", Proceedings of the 2017 International Conference on Arabic Script Analysis and Recognition (ASAR), IEEE, 2017
4. Darshana Patil et al., "Novel Technique for Script Translation using NLP: Performance Evaluation," Proceedings of the 2021 International Conference on Natural Language Processing (ICNLP), IEEE, 2021
5. Hritik Singh et al., "Multilingual Education through Optical Character Recognition (OCR) and AI", Proceedings of the 2020 International Conference on Education and Artificial Intelligence (ICEAI), IEEE, 2024
6. Manish Kumar Gupta et al., "Chitrantaran: Web-based Platform to Enhance the Document Digitization Process using OCR and Machine Translation", Proceedings of the 2024 International Conference on Computational Linguistics and Intelligent Systems (CLIS), IEEE, 2024
7. Kolla Bhanu Prakash et al., "Text Studies towards Multi-lingual Content Mining for Web Communication", Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2010
8. Kolla Bhanu Prakash et al., "Information Extraction in Unstructured Multilingual Web Documents ", Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2015

9. Kolla Bhanu Prakash et al., "ANN for Multi-lingual Regional Web Communication", Proceedings of the 2012 International Conference on Soft Computing and Pattern Recognition (SoCPaR), Springer, 2012
10. Hruday Shetty et al., "ROLEX: AI-DRIVEN VOICE-CONTROLLED ROBOTIC SYSTEMS: ENHANCING HUMAN-ROBOT INTERACTION ", International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS), 2025
11. Balaji D et al., "VIRTUAL PERSONAL AI ASSISTANT: ENHANCING TASK AUTOMATION WITH GENERATIVE AI AND NLP ", International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS), 2025
12. Gongbo Tang et al., "Improving BERTScore for Machine Translation Evaluation through Contrastive Learning ", Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), IEEE, 2024
13. Rohitash Chandra et al., "Semantic and Sentiment Analysis of Selected Bhagavad Gita Translations Using BERT-Based Language Framework ", Proceedings of the 2022 International Conference on Language Resources and Evaluation (LREC), IEEE, 2022
14. Ying Fang et al., "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness", Proceedings of the 2018 International Conference on Information Systems and Design of Communication (ISDOC), IEEE, 2018
15. Tatiana Sherstinova et al., "Sentiment Analysis of Literary Texts vs. Reader's Emotional Responses ", IEEE Xplore Digital Library, Proceedings of the 2023 International Conference on Computational Linguistics (COLING), 2023
16. Gullilermo Marco et al., "Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches ", IEEE Access, 2021
17. R Chandra et al., "Large Language Models for Metaphor Detection: Bhagavad Gita and Sermon on the Mount", IEEE Access, 2024
18. M Dusi et al., "Discrimination Bias Detection through Categorical Association in Pre-Trained Language Models ", IEEE Access, 2024
19. Simona Tiribelli et al., "Ethics of Artificial Intelligence for Cultural Heritage: Opportunities and Challenges", IEEE transactions on Technology and Society, 2024
20. A Romanov et al., "Integrated Technique of Natural Language Texts and Source Codes Authorship Verification in the Academic Environment ", IEEE Access, 2025

Future Scope:

The future scope of this project includes:

1. **Support more ancient and low-resource languages:** Expanding the language library will allow the tool to handle rare scripts and historically important manuscripts that currently lack automated processing.
2. **Implement transformer-based OCR and restoration models:** Using advanced architectures like Vision Transformers and Seq2Seq Transformers will improve accuracy when processing heavily damaged or unclear text.
3. **Add manuscript image processing:** Allowing users to upload scanned pages or inscriptions will enable end-to-end analysis from raw images to restored text and meaning.

4. **Enhance linguistic analysis depth:** Including detailed morphology, phonology, syntax, and dialect variations will provide richer philological insights for researchers.
5. **Develop language-specific rule engines:** Tailored rules for each language can improve correction accuracy and produce more authentic, culturally consistent outputs.
6. **Introduce historical metadata prediction:** The system can be upgraded to estimate author style, era, region, or period, helping researchers understand the historical context of texts.
7. **Continuously retrain models with larger corpora:** Training on more diverse datasets will help the tool adapt to new languages, writing styles, and historical variations.