



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Philological Analysis Tool Using AI

¹Anjali.K.S, ²C D Subramanya Iyengar, ³Dhanush R, ⁴Dr.M.A. Dorai Ranaga swamy

¹⁻³ Final year Student, Department of AIML, Bangalore Technological Institute, Bangalore ⁴Senior Professor, Department of AIML, Bangalore Technological Institute, Bangalore ¹Bangalore Technological Institute, Bangalore, India.

Abstract: This paper presents a comprehensive Philological analysis tool using AI designed to assist the learners, people who want to gain knowledge of Sanskrit and Proto Dravidian languages. Language is the base of communication and for gaining knowledge. Knowledge that has been used by our ancestors and to know the different concepts, and techniques that were used for solving a particular problem. To see about Sanskrit as we know that it was the language that was mostly used for documentation of the ideas as well as to transfer knowledge from generation to generation over the period of time. One of the speciality is that the language is versatile and is simple to use in that period of time. As the time passed, many languages have evolved from Sanskrit and have given foundation for the present languages in use. The information was stored in the palm leaves and other format as years pass they will be damaged. The another language that we will be dealing is Proto Dravidian language that was basically used in 2500BC around this time. The Proto Dravidian language is an old language that was used in the ancient times and in the present world we can't find the speakers of this language. Due to the gap in the time period. From this language the south Indian languages has evolved over time like Telugu, Tamil, etc. Now, to look back to the origin language will be a huge task for collecting the data as well as to understand it and will be challenging to work on this project so that we can explore the concepts and ideas of the past. In this project we are tried to use the tools of AI in which NLP(Natural Language Processing) is the correct concept with other sub fields of AI. AI is booming in this era and one of the important field for language translation. Translation of Sanskrit and Proto Dravidian languages to English is the main task will be done in this project work for the users to know the meaning and concepts of the past and gain insights of the culture. AI is very much helpful in performing task based on training data and testing data to evaluate the model of the project. This project is helpful when people want to know the information without explicitly learning the language. The process of learning language is not taken as a priority in the work because many people may or may not know the language and gaining knowledge is for everyone and should not be restricted by language boundary. For new learners and people who want to gain knowledge without learning the language can rely on AI for this task of language translation.

Index Terms – Sanskrit, Artificial Intelligence, Natural Language Processing, Optical Character Recognition, Proto Dravidian language

I. INTRODUCTION

Language is the base for communication between people over the years as well as now. Language plays an important role in making people understand and respond with each other and to express their thoughts and experience. This project we are concentrating on Sanskrit and Proto Dravidian languages for translating them into English so, that do not know the language can also understand the historical ideas and concepts that were in use. Here, the user need not learn the language to understand the things but the NLP module will perform the repetitive process of translation . We are mainly using OCR(Optical Character Recognition), NER(Named Entity Recognition) to do this task of language translation in our project

1.1 PROBLEM STATEMENT

In this project we are taking the input in the form of text based in Sanskrit or Proto Dravidian language then, it will be understood as well as interpreted for further information extraction from the input. The main process involved in this project is to translate Sanskrit and Proto Dravidian languages into the English language with the help of AI and it's concepts. AI is booming in all the industries in the recent years and also very helpful in doing the repetitive tasks. As the process of translation needs experts of particular language. We may get experts in the Sanskrit language but for Proto Dravidian languages is very very difficult as it was the language that was used so many years ago. And also so much of work has to be done in manual process and also time consuming with the help of AI the process can simplified and be more accurate with proper training and testing process.

1.2 OBJECTIVE OF THE PROJECT

Language is a form of knowledge transfer that is used over years to transfer information from one generation to another. Knowledge transfer is very important to know the roots and concepts of our ancestors that was already in practice and use them for further development of ideas in the present.. Sanskrit and Proto Dravidian languages both are transformed or divided into various languages in the present world. So, to get back to the root language will be difficult and challenging to find the information related to these languages. In the present world Sanskrit is divided into Hindi, Bengali, Marathi, Gujarati which are part of Indo-aryan languages. Similarly if we take Proto Dravidian language in the present it is divided into Tamil, Telugu, Kannada, and Malayalam.

The primary objective of this project is to design, develop and implement an AI powered Philological analysis tool that is capable of digitizing and interpreting ancient manuscripts written in classical Indian scripts such as Sanskrit and Proto dravidian languages. We are taking the input as Sanskrit or Proto Dravidian language and will pass through the process of understanding and interpretation of the information present in the input. This project is generally talking the information or input from Sanskrit and Proto Dravidian languages and the model will be trained on the dataset and also able to predict the language of the input. We are processing the input to give the required translated information of the input is the main task performed here. The project

will be helpful in understanding the language that is Sanskrit and Proto Dravidian languages and also interpret and knowledge the thoughts as well as the ideology of the people during that period.

1.3 SCOPE

- To integrate large Datasets with more parameters.
- The use of advanced techniques for making the process to be simplified.
- Improving the language translation efficiency and accuracy.
- To integrate more languages for the output or outcome.
- To integrate it to an interactive website or webpage.
- To integrate chatbot for more user friendliness.

1.4 PROJECT INTRODUCTION

To begin with philology which is an interdisciplinary study of language in the historical and literary contexts that plays a foundational role in the process of preserving the intellectual and cultural heritage of civilization. This is mainly deep rooted in the detailed analysis of the ancient texts, literature, philosophy, history and linguistics. The linguistic traditions are not merely vehicles of communication but are also the repositories of ancient knowledge, religious thoughts, religious ideology, scientific enquiry and cultural expression that have shaped Indian continent.

The traditional Philological research have faced a lot of problems and challenges and also takes lot of efforts yet the hard work pays off. Mainly we have to depend on the rare, fragile manuscripts that are difficult to access and as well as difficult to understand them as language have evolved over the duration of time. In the recent years there is a lot transformation in this field because of AI as it has opened new possibilities for Philological analysis. There is also increase in the digitalization of the manuscripts, combined with the powerful concepts of AI models, enables a shift from manual and isolated efforts to the scalable and collaborative work. Some of the major tools are Optical Character Recognition (OCR), Natural Language Processing (NLP), Machine translation and Deep learning.

Sanskrit is one of the oldest attested Indo-European language of the world that has vast and also diverse literary corpus ranging from religious scriptures like the Vedas and Upanishads to epics like the Ramayana and Mahabharata which are the treasures for the followers. In parallel to see about Proto Dravidian language which is reconstructed language that represents the ancestral root of modern Dravidian languages such as Tamil, Telugu, Kannada and Malayalam. By giving importance to linguistic traditions and the practical difficulties in studying them manually that is time consuming and needs lot of efforts.

The foundation of the project begins with the advanced OCR engine that is fine tuned for region specific language scripts like Devanagari. The digitalization of documents is only the first step that we have to take for Philological analysis as storage of information plays a key role in the process of extraction and analysis of data. To extract the meaningful information present in the manuscripts will be extracted and interpreted to English language. There are many of transformer models of AI that are helpful in this process.

This project will show the convergence of AI and philology which is not only the Technological upgrade further it will also mark the shift in the way we study, understand and also interact with only the historical insights. We are making a stand such that as a step forward towards a future where by digitalizing, analysing and interpreting ancient texts at scale. We can also unlock the opportunities to not only historical insights but also cultural wisdom that continues to resonate.

2. Literature Review

Table.2.1 Comparitive study table of three researched paper

Criteria	Paper 1	Paper 2	Paper 3
Title	CNN- LSTM based approach Recognition of Devanagari Manuscripts.	Deep Learning based approach for historical manuscript dating.	A database for Degraded Arabic Historical Manuscripts..
Year of Publication	2022	2019	2016
Motive	To recognize and return Devanagari text in historical manuscript using a deep learning pipeline that combines facial and sequential processing.	To automate the dating of historical manuscripts using deep learning, enabling accurate temporal classification of undated and uncertain documents.	To provide a benchmark data set of degraded Arabic manuscript to support research in document analysis and restoration.
Data collection	Devanagari manuscript data sets comprising real historical text and augmented synthetic handwritten samples.	Historical manuscript images with confined dating metadata collected from digital our chief and historical repositories.	Real historical Arabic manuscript images with degradation such as ink bleed stains and feeding from archival sources.
Pre processing	Grayscale, conversion, normalisation, resizing and background noise removal to enhance image quality and text clarity.	Images undergo normalisation, resizing and noise removal data is augmented to increase robustness and decrease over fitting.	Binarization, noise removal, skew correction, and image enhancement apply to scan manuscript images.
Models Used	Hybrid CNN -LSTM architecture where the CNN extract spatial text features and the LSTM handles the sequence predictions enabling accurate recognition of hand written script.	A Convolutional Neural Network (CNN) architecture is trained to learn temporal features and patterns from visual man script to classify documents by era.visual man	No specific model proposed the data set provided will facilitate the algorithm development and benchmarking.
Results &	It achieved highest accuracy of 93.63 percentage with a 75: 25 training to testing ratio at 27 epochs. The finding	InceptionResNetV2 achieve the best performance with a document level m ae of 3.0 1 years and a cumulative	Sauvola's and Gatos' methods performed best in handling various degradation types. Global thresholding

	suggest that increased training data and training time improve the model's performance.	accuracy of 98.101% within a plus aur minus 25 years margin.	(Otsu) short poor performance on low contrast and bleed through images.
--	---	--	---

3. *SYSTEM ANALYSIS*

3.1 *Introduction:*

The process of developing analysis tools for ancient manuscripts demand for a multi disciplinary approach where in we need to take many of the components such as computational linguistics, deep learning and historical language research which are the main components. To see about the overview of the system analysis that we are going to use in this project will be discussed in the further part of the paper. All together the components of the system will be well defined sustainable and aligned with the real world applications in philology and man script preservation so that it will be beneficial for others also in the future to extract the information from the manuscripts.

3.2 *System requirements:*

- Optical character recognition (OCR):
Near the core function of the tool will be taken care of like taking the high resolutions can't images of the manuscript and extracting the text from the images. The OCR will be responsible for handling variations in the script style ink, degradation, faded text and uneve lightning.
- Script and language identification:
In this module when we upload an image the system should automatically detect the language and the script type. This will also enable the OCR engine to lower the appropriate model configuration for specific scripts of each language.
- Text translation:
After the process of recognizing the text it must be translated to English. This process will include preserving semantic richness, synthetic structure and contextual information. Yeah fine tune model trained on parallel corpora of classical text can be used for this process.
- Authorship attribution:
In the process of authorship attribution we are going to majorly analyse linguistic cues of the author such as syntax, diction, stylistic marker and the style should infer likely authors or schools of other ship. This can be facilitated through machine learning classifiers friend on known copper and stylometric features.
- Interactive user interface: The present should be visually appealing to the user so that it will be very easy for the user to understand the interface and do the task whatever they want to perform. Will include modules for uploading images displaying OCR and translation results browsing metadata and interaction with an AI assistant that can guide the uses in exploring various features.

4. *ARCHITECTURE*

The architecture for this AI best philological analysis tool is modular and layer system engineer to address the multifaceted challenges of interpreting ancient and degraded manuscript text to English. Manuscripts will often suffer from physical deterioration, complex script structures, deep semantic richness, which are helpful for the integration of advanced image, character recognition and linguistic analysis. The design of this architecture of the project emphasis on scalability, adaptability to multiple scripts, and the usability for researchers and scholars who are interested in the historical text analysis process. The system include several layers that work in a sequential pipeline pipeline.

The input layer will handle manuscript uploads in high resolution image format or text can also be taken for the process of translating. Next the pre processing layer cleans all the images and text by reducing noise, enhancing contrast, and by applying binarization for distinguishing text . It will also segment the images into smaller patches for more accurate processing of the information.

The NLP and translation module will take care of analysing the linguistic features of the given text. It will perform the morphological analysis and part of speech tagging as well as name density recognition for the process of translating the given language text into English by using models. All the models in the architecture will work collectively with each other to give the required output or outcome.

The Named Entity Recognition (NER) module will be dividing the given text into different categories and the naming it under the specified category so that it will be easy in the future to retrieve any information whenever required. This is one of the technique that is used for dividing the text into different categories and making it easy for information retrieval process as we can search based on the category and we can get the information easily.

The project as various modules in the architecture which will be dealing with different process that has to be taken place for translating Sanskrit and Proto Dravidian languages to English by understanding the given input interpreting them and going to all the modules and completing the process to get the required output. Modules play a important role in this process as it include several components in this project and each of the module will act individually and contribute to the efficient and accurate prediction of the language and giving the appropriate translation for the given text.

The process of translation corpora that is available for equivalent English translation available for Sanskrit we have wide range of data with corresponding translation given to it in English. When we take Proto Dravidian language the data set is scattered and need to be joined together to form efficient data set on which the model can be trained. The process of evaluation of the model is also important as we have to check the accuracy and our efficient the model is when translating the input to the required output. Testing place a major role in this project for knowing whether the translation made by the model is accurate or not. The architecture can be seen in fig 4.1.

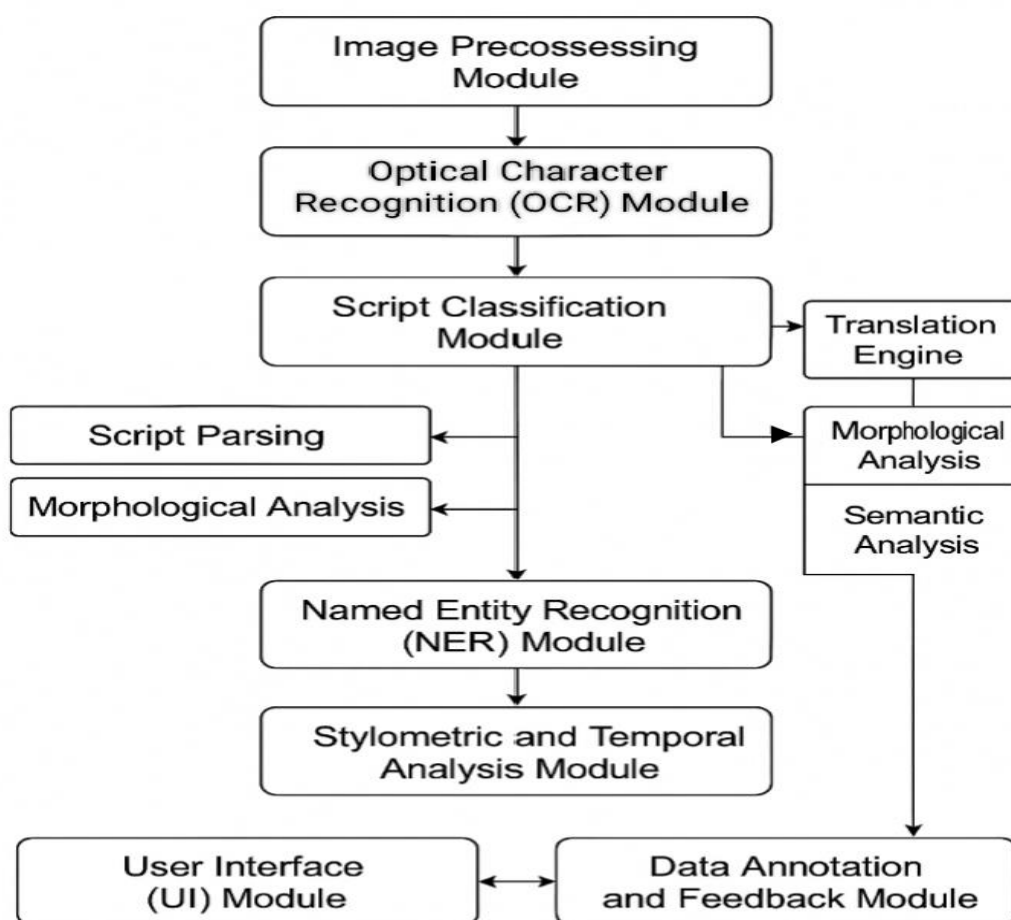


Fig 4.1 Architecture diagram

5. Conclusion

The development of this project which is 'A comprehensive Philological analysis tool using AI' mainly represents the emerging of Artificial Intelligence with classical language preservation and analysis how AI can help in understanding and interpreting as well as transferring knowledge from one generation to another.

The AI is the most evolving field that is getting recognised in other fields acting as an interdisciplinary field in the industry. The use of ancient man is scripts which are often degraded, fragmentary, linguistically complex force significant challenges for historians and philologists aur analysing and translating it to the English. By employing a modular architecture integrating cutting edge technology such as image pre processing, OCR, script classification, NLP, machine translation and metadata extraction.

Each module of the project is designed to function both independently and as a part of collaborative pipeline which will enable robustness, performance across a variety of scripts and manuscript conditions. A user friendly interface is very important for developers to attract more and more users. This system can enhance usability and promotes collaborative refinement. As development progresses in two model training and interface deployment the model will be able to support academic research and educational use and digital preservation at scale so that many uses can use the information and also gain knowledge.

Ultimately this project AIIMS to bridge the historical knowledge gap and make access tool linguistic heritage and poster ready for understanding of the classical text. It is not nearly a technical solution but so a digital garden of cultural memory preserving the past with the intelligence of the future.

REFERENCES

- [1] Aditi Moudgil et al.,(2022) CNN-LSTM Based Approach for Recognition of Devanagari Manuscripts, IEEE xplore-proceedings of the 2022 International conference on Data analytics for business and industry.
- [2] Hamid et al.,(2019),Deep Learning Based Approach for Dating Historical Manuscripts: a study of CNNs Applied to Paleographic Analysis, IEEE xplore -Proceedings of the 2019 international conference on document analysis and recognition.
- [3] Sulaiman et al.,(2017), A Database for degraded Arabic Historical Manuscripts, IEEE xplore- research founded by FRGS/1/2016/ICT02/UKM/01/1.
- [4] Darshana Patil et al.,(2021), Novel Technique for Script Translation using NLP: Performance Evaluation, IEEE xplore -Proceedings of the 2021 International conference on emerging smart computing and informatics, AISSMS Institute of information technology.
- [5] Hritik Singh et al.,(2024), Multilingual education through Optical Character Recognition (OCR) and AI, IEEE xplore -Proceedings of the MIT Art, designer technology School of computing international conference.
- [6] Manish Kumar Gupta et al., (2024), Chirantaran:Web based platform to enhance the document digitalisation process using OCR and Machine translation, IEEE xplore -Proceedings of the 4th disciplinary conference on electrics and computer.
- [7] Hruday Shetty et al.,(2025),ROLEX: AI-DRIVEN VOICE-CONTROLLED ROBOTIC SYSTEMS: ENHANCING HUMAN-ROBOT INTERACTION, International Research Journal of Modernization in Engineering Technology and Science.
- [8]Balaji D et al.,(2025),VIRTUAL PERSONAL AI ASSISTANT: ENHANCING TASK AUTOMATION WITH GENERATIVE AI AND NLP, International Research Journal of Modernization in Engineering Technology and Science