INTEGRATING MACHINE LEARNING AND NLP EFFICIENT RETRIEVAL OF CHARACTERS IN PALI SCRIPT PRESERVATION

Sangita R. Gudadhe, Dr. Aashish A. Bardekar, Department of Computer Science & Engineering, Sipna College of Engineering Technology,Amravati,Maharashtra,India(444602)
Dr. Amitkumar B. Ranit, Department of Civil Engineering, Prof Ram Meghe College of Engineering and Management, Amravati. amit.ranit@gmail.com

Abstract:
Pali is an ancient as well as Prakrit language having enormous data of knowledge and literature but Pali language consider as a religious language. Because of this consideration many of the people are not aware about the huge knowledge of this literature. Pali language is not a religious language which is used only to express the thoughts of Buddha because Pali was the people's linguistic language at the time. Historical and medical data also included in Pali linguistic literature. The aim of this proposed research work is to retrieve and extract information from literature, differentiate between the meaning of Pali words and convert to match proper synonym words. This is for those people who are inspired by ancient text which is actually written in Pali literature and interested to learn, acquire knowledge,Also help to preserve our traditional and ancient literature by using recent tools and technologies.Using Optical Character Recognition (OCR) technology we will first recognize image and modified distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document.Later using Natural Language Processing (NLP) combines the field of linguistics and computer science to decipher Pali prakrit language structure and grammatical phrases to make models which can comprehend, break down and separate significant details from scripted literature.

Keywords - Ancient scripts, Meaning, Pali, OCR, Natural Language Processing.

1. Introduction

Language is a tool to express our inner thoughts, feelings, thoughts etc. Man has been fascinated by language since ancient times. There have been various legends and stories about the origin of language and power in folk tradition since very early times, and this is an attempt to satisfy that curiosity. In our India many more are the ancient languages. Among that languages Pali is one of the ancient languages having a sweetness and soberness of nature. Because Pali is a Prakrit language. Pali language is also called as a classical and liturgical language of the Theravada Buddhist.

To provide naturality and purity to this Pali language, we are using Pali language for research because everybody can take advantage of this language.

Our aim is to provide the meaning full impact of this language to everyone. So, everybody go through this language and take a advantages of this language. This is also beneficial to those people who are inserted in Vipassana. Because a whole description of Vipassana 's included only in Pali language.

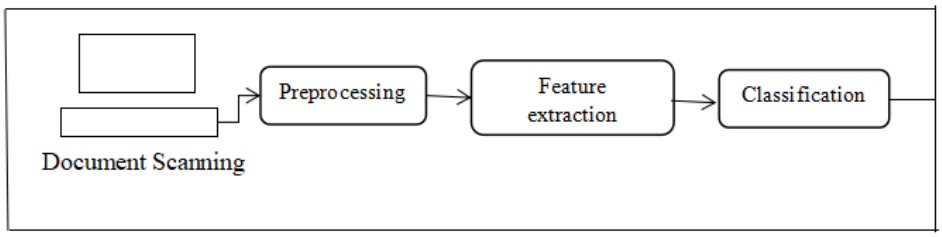


Figure 1: Components of OCR System

Humans and machines are incompatible without each other. Handwriting recognition is the ability of any machine to identify and interpret handwritten input. This can be done online or offline. Manuscripts written in ancient times are the only mark of their civilization and should be preserved efficiently. This early writing was called cuneiform and consisted of making specific marks on wet clay using reed tools.

Optical Character Recognition (OCR) is a technology that forms the roots of image recognition and has been modified many times over many functions. Here we use his modified OCR to recognize the Aryaka script used in the 4th century BC. Used to write the Pali language since around 1000 BC, it helps archaeologists understand the rich history of the Pali language and the thoughts, culture and teachings of the Buddha. Pali has 41 alphabets in total, with 32 consonants, 6 vowels, 2 diphthongs, and one accessory nasal sound called Niggah'ta [14].

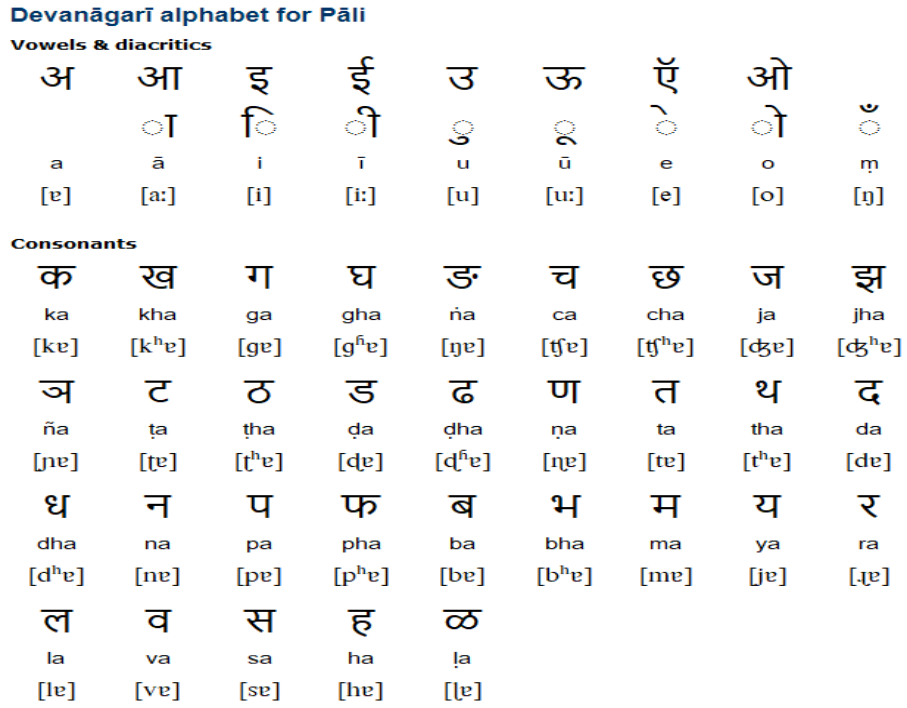


Figure 2: Devanagari alphabet for Pali language

Figure 3: Devanagari alphabet for Pali language with examples

Nouns are inflected for gender, number, and case; verbal inflections convey information about a person, number, tense, and mood. Pali nouns inflect for three grammatical genders (masculine, feminine and neuter) and two numbers (Singular and plural). The nouns also, in principle, display nine cases. Pali language is also called as the language of the scriptures of Theravada Buddhism that

is nothing but the Tipitaka which were written in Sri Lanka during the 1st century BC. Tipitaka is also written in Pali language. Pali has been written in a variety of scripts, including Brahmi, Devanagari and other Indic scripts.

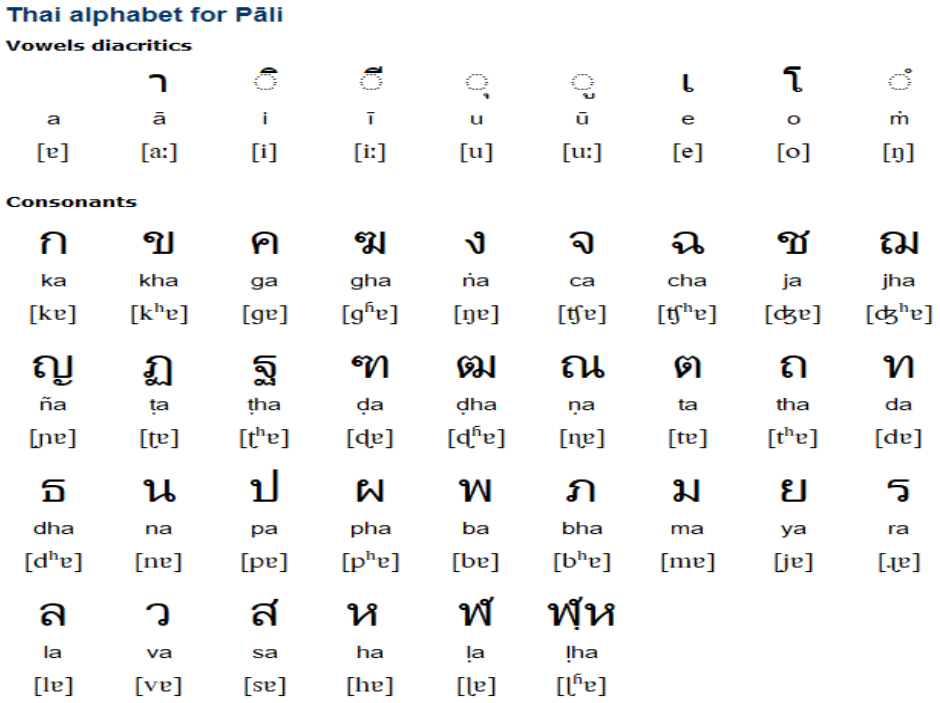
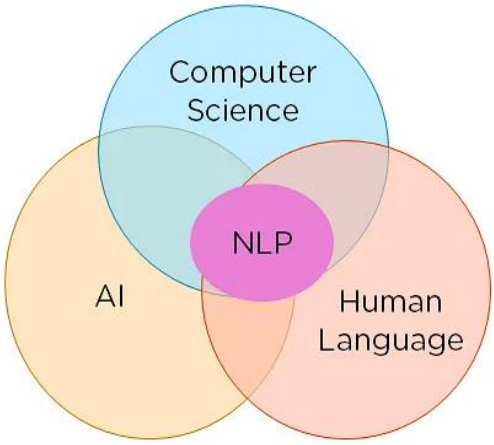


Figure 4: Thai alphabet for Pali language

1.1 Role of Natural Language Processing (NLP) -

Humans communicate with each other using words and text. The way that humans convey information to each other is called Natural Language. Every day humans share a large quality of information with each other in various languages as speech or text. However, computers cannot interpret this data, which is in natural language, as they communicate in 1s and 0s. The data produced is precious and can offer valuable insights. Hence, you need computers to be able to understand, emulate and respond intelligently to human speech.

Natural Language Processing (NLP) combines the field of linguistics and computer science to decipher language structure and guidelines and to make models which can comprehend, break down and separate significant details from text and speech.



There are some steps to perform to perform the operation of Natural Language Processing (NLP). The steps to perform preprocessing of data in Natural Language Processing (NLP) are as follows:

Segmentation: For segmentation need to break the document into its constituent sentences. A document can break with its punctuations like full stops and commas.

ISSN: 2278-4632
Vol-14, Issue-4, No.01, April: 2024

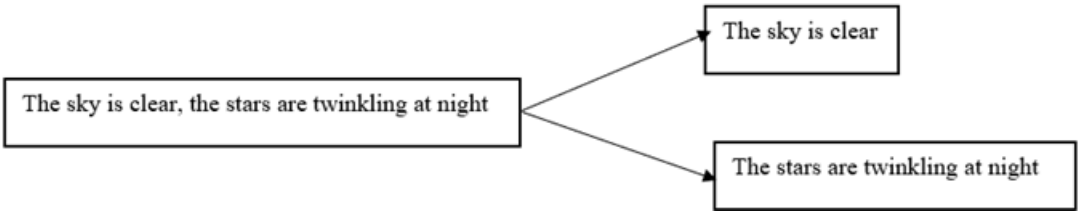


Figure 5: Segmentation

To kenizing: To get the words in a sentence and explain them individually to our algorithm, you must first get the words in a sentence. So we can divide our sentence into constituent words and save them. This is known as tokenizing, and each world is referred to as a token.

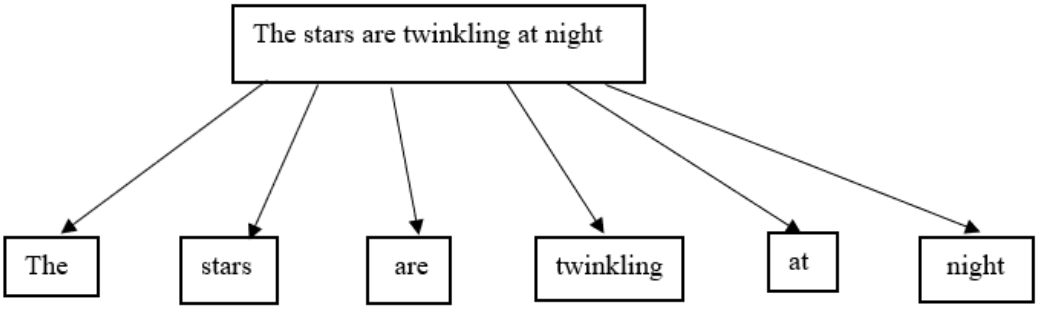


Figure 6: Tokenization

Removing Stop Words: To speed up the learning process by removing non-essential words that add little meaning to our statement and are only there to make it sound more cohesive. Stop words are words that can be removed, such as was, in, is, and the.

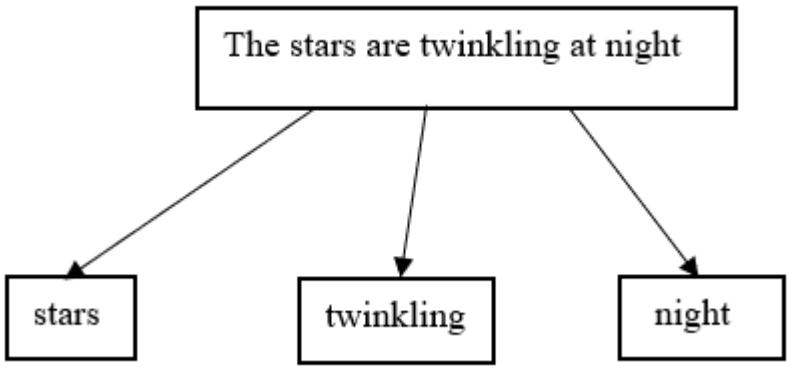


Figure 7: Stop Words

Stemming: Stemming is the process of obtaining the Word Stem of a word. Word Stem gives new words upon adding affixes to them

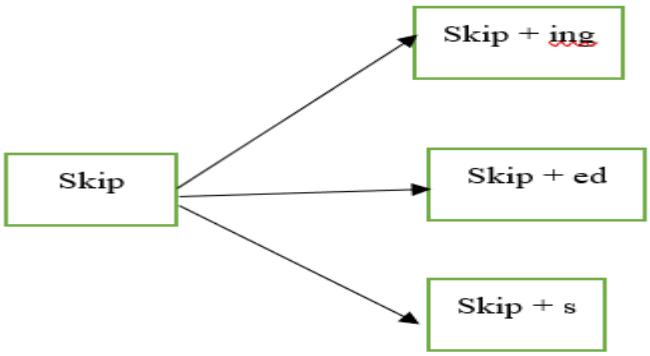


Figure 8: Stemming

Lemmatization: The process of obtaining a word's Root Stem. Root Stem provides the new base form of a dictionary word from which the word is derived. To identify the base words for various words based on their tense, mood, gender, and so on.

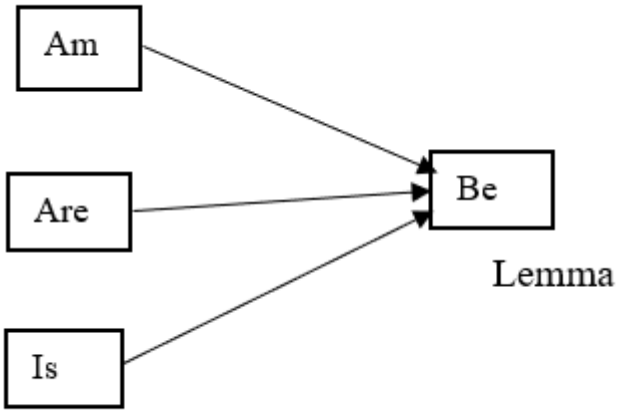


Figure 9: Lemmatization

Part of Speech Tagging: Now explain to the machine the concept of nouns, verbs, articles, and other parts of speech by adding these tags to our words. This is known as 'part of'.

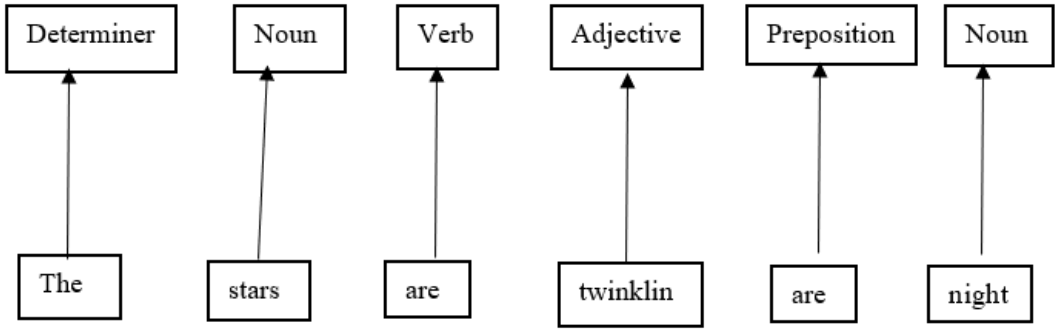


Figure 10: Part of Speech Tagging

Named Entity Tagging: Next, introduce our machine to pop culture references and common names by flagging movie titles, important personalities or locations, and so on that may appear in the document. This is accomplished by categorizing the words into subcategories. This allows us to locate any keywords in a sentence. Person, location, monetary value, quantity, organization, and movie are the subcategories.

There are multiple applications of Natural Language Processing like as

1. Question Answering (Alexa)-Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.
2. Spam Detection - is used to detect unwanted e-mails getting to a user's inbox.
3. Sentiment Analysis - is also known as opinion mining. It is used on the web to analyze the attitude, behavior, and emotional state of the sender.
4. Machine Translation - is used to translate text or speech from one natural language to another natural language.

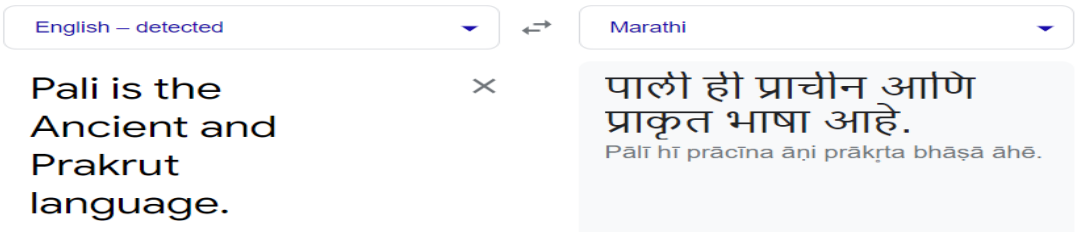


Figure 11: Translated data

5. Spelling correction-Microsoft Corporation provides word processor software like MS-word, PowerPoint for the spelling correction.
6. Speech Recognition - is used for converting spoken words into text. It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.
7. Chatbot - Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.
8. Information extraction - is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.
9. Natural Language Understanding (NLU) - It converts a large set of text into more formal representations such as first-order logic structures that are easier for the computer programs to manipulate notations of the natural language processing.

2. Literature Review

A lot of work has been done by OCR in various languages and scripts, providing high accuracy and leaving the standard to others. In 2004, U. Pal and B. B. Chaudhari [3] reported a large OCR work on 12 Indian characters.OCR is identification of text, which may be printed or hand-written[4]. 2007 V.N. Manjunath Aradhya, G. Hemantha Kumar, S. Noushath [5] demonstrates highly accurate multilingual OCR. Later, Apurva A. Desai [6] 2010 used his OCR technology to recognize handwritten Gujarati digits with a success rate of 82. Amit Choudhary, Rahul Mishra, and Savita Ahlawat [7] also routinely do the same with binarization methods, and his ability to evaluate OCR of English letters in 2013 shows an accuracy of 85.62% . Pali character recognition using Devnagri was designed in 2012 by Kiran S. Mantri, S. P. Ramteke, and SR Suralkar [8], incorporating features such as image preprocessing, feature extraction, and classification algorithms. Proven by converting to software (OCR). High probability production. A 100% detection rate was achieved using a simple feedforward multi-layer perceptron, and we also propose a backpropagation learning algorithm used to guide each network using a specific group of characters. Another study recommending a recognition system incorporating the Pali map of Buddha Dasa Indapano was published in 2013 by Thanathany Fientrakland and Wangwisa Chevakurmongkol [9]. His hand-drawn images were refined through contrast adjustment, grayscale conversion, and noise removal. Basically, individual character features are removed by the zonal method, and the average accuracy of all considered groups is about 81.73%. Machine translation is a subfield of natural language processing identified in early artificial intelligence (AI)[11]. Machine learning and deep Neural Network techniques have been widely used for automatic recognition of characters and digits of different languages, and in various classification-based problems [12].The process of machine translation can be simply described as decoding the meaning of the source text and re-encoding that meaning into the target language. However, due to the complexity of natural language, the development of machine translation systems has become a research topic. Machine translation systems can be broadly classified into direct translation systems and indirect translation systems according to the translation method. Direct translation systems use word-to-word or phrase-to-phrase mappings to translate a source language into a target language. Indirect translation systems use an interlingua or transfer approach. In addition to the common machine translation approaches mentioned above, machine translation systems can be classified into seven categories. human-assisted, rule-based, statistical, example-based, knowledge-based, hybrid, and agent-based [15]. Dictionary-based translation is used as a basic approach to machine translation. This type of machine translation application is easy to implement and uses very capable tools to get translations for unknown search terms. Usually, dictionary-based translation is essentially translation using a bilingual dictionary. In addition to bilingual dictionaries, most systems consist of a morphological analyzer and source and target language generators. For example, Sindhu et al. developed a dictionary-based machine translation from Kannada to Telugu [16]. The translation system consists of five components: Morph Analyzer, Dictionary, Transliteration, Conversion Grammar and Morph Generator. Fields of text mining and

natural language processing provide tools and methodologies that can help with retrieving relevant information [17].
Consider the Pali language, there has been very little research into translating Pali text into other languages. Phoson and colleagues have created a rule-based machine translation prototype for Pali to Thai [18]. This system can analyse the Pali language structure of an input sentence and generate Thai language structure.

3. Proposed Approach of Methodology

The classical paradigm for character recognition has three steps: segmentation, feature extraction, and classification. In segmentation we will attempt to segment words into letters or other units using or without feature-based dissection algorithms. A vertical scan is the most basic and straightforward segmentation algorithm. The algorithm divides the image into black and white pixels and then searches for unbroken columns of white pixels. This works well for machine-printed or handwritten characters where a certain amount of white space is required. Isolating regions of connected black pixels is a more robust technique. This method divides the black pixels into sets, with each black pixel in the set adjacent to another black pixel. This method works exceptionally well for digits that are not overlapping, touching, or disjoint.

A more compact and characteristic representation is required in most recognition systems in order to avoid extra complexity and increase the accuracy of the algorithms. As a result, we will try the following methods for feature extraction, i.e. representation:
Neural Network - A neural network is a computing architecture made up of massively parallel interconnections of adaptive 'neural' processors. Because of its parallel nature, it can perform computations at a faster rate than traditional techniques. It can adapt to changes in the data and learn the characteristics of the input signal due to its adaptive nature. A neural network is made up of many nodes. The output of one node in the network is fed to another in the network, and the final decision is determined by the complex interaction of all nodes.Convolutional Neural Networks (CNNs): CNNs are deep learning models that have been shown to be extremely effective in image recognition tasks such as OCR. CNNs learn to detect image features relevant to character recognition and classify the characters based on those features.Support Vector Machines (SVMs): SVMs are supervised learning algorithms that can be applied to OCR. They are especially good at recognizing characters with well-defined features, such as printed text.These are some of the AI-based OCR methods. Depending on the application and the characteristics of the input data, there are numerous other techniques and algorithms that can be used.

3.1 Proposed Work

Optical character recognition consists of variety of components with this the input images are distributed in two form ie. are training and testing images. Optical Character Recognition (OCR) is a technology that forms the roots of image recognition and has been modified many times over many functions for getting the maximum accuracy of the output

ISSN: 2278-4632
Vol-14, Issue-4, No.01, April: 2024

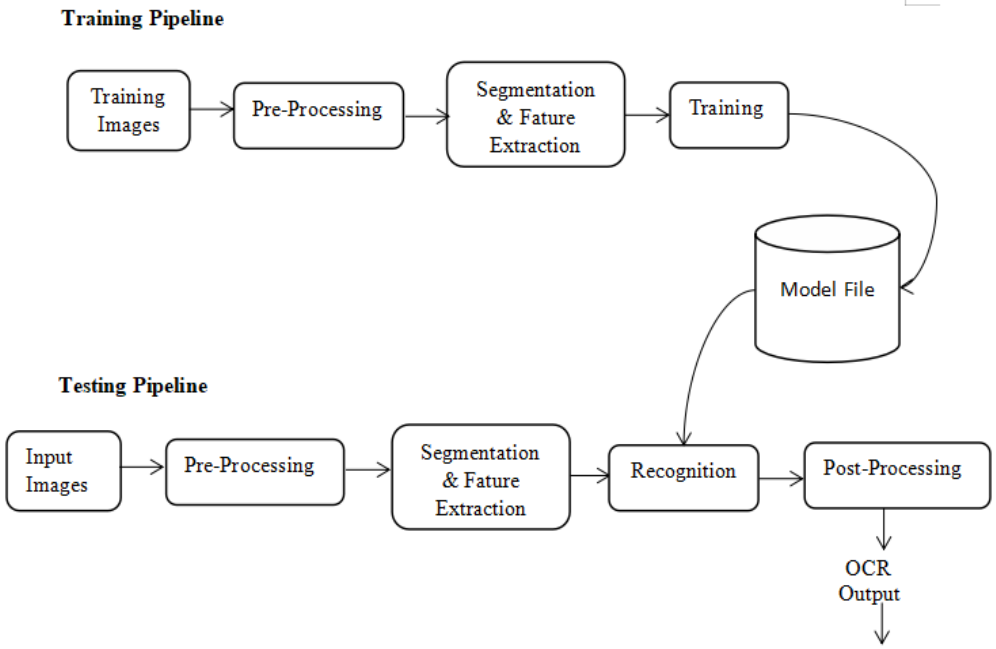


Figure l2: Proposed work Diagram

6. Result Analysis

The complete process as discussed above was performed on the data set used. Hence the training of the system was done by this way. The testing was performed by giving the printed Pali script in two types : first on all the individual letters and then secondly on the sentences. For the letters the accuracy of individual character recognition is 92% while working on the sentence the accuracy decreased
by few count.

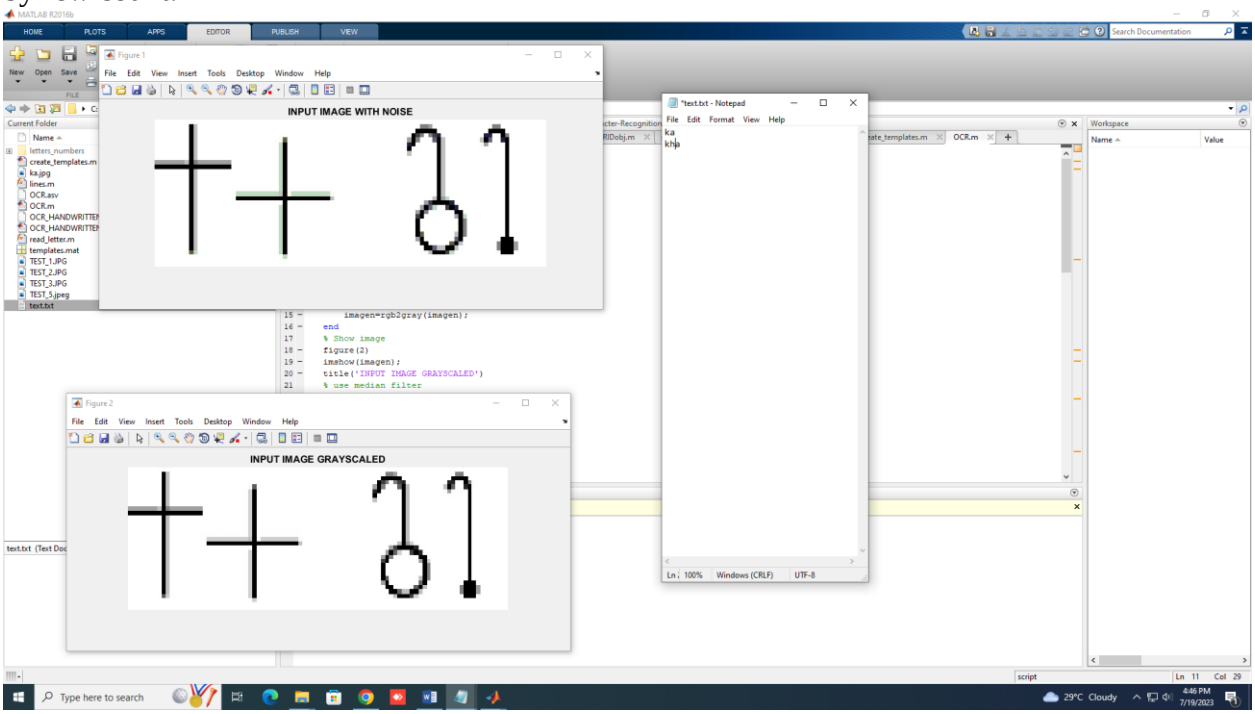


Figure 13: Result of individual letters

Sr.No.	No. Sample	No of Classifications	No. of Misclassifications	Accuracy	Time (Sec.)
1	25	23	2	92%	1.26

Table 1: Classification table for individual letters

7. Conclusion

Today, Pali language is studied mainly to gain knowledge from ancient literature. Because the extraction and retrieval of such rich ancient Indian literature humongous information contributes to the opening of doors to innovators and that modern languages cannot provide. Using NLP, we can retrieve and extract information efficiently from literature, distinguish between the meanings of Pali words, and accurately convert to match the right synonyms.OCR will help to preserve and distinguish printed or handwritten text characters of our traditional and old literature.

References

[1] Vijayalakshmi R., Dr. J. M. Gnanasekar "A Review on Character Recognition and Information Retrieval from Ancient Inscriptions", 2022 8th International Conference on Smart Structures and Systems (ICSSS) | 978-1-6654-9761-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSSS54381.2022.9782241

[2] An Elementary Pali Course book by Ven Narada, Thera pg. no. 9-12

[3] U. Pal B.B. Chaudhari, "Indian scripts character recognition: A Survey Pattern Recognition Society", Published by Elsevier Ltd 2004

[4] Preetha Sa, Afrid I Mb, Karthik Hebbar Pc, Nishchay S Kd, "Machine Learning for Handwriting Recognition", International Journal of Computer (IJC) ISSN 2307-4523.

[5] V.N.M Manjunath Aradhya , G. Hemantha Kumar, S. Noushath, Multilingual
OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis Engineering Applications
Artificial Intelligence Elsevier Ltd 2007.

[6] Apurva A. Desai, "Gujrati handwritten OCR through neural network Pattern Recognition" , Elsevier Ltd. 2010

[7] Amit Choudhary, Rahul Mishra, Savita Ahlawat, "Off- line handwritten character recognition using feature extracted from binarization technique", The Authors. Published by Elsevier B.V 2013

[8] Kiran S. Mantri, R. S. Ramteke,,"Pali Character Recognition System: A Survey", at IJAIR 2012 ISSN: 2278-7844

[9] Tanasanee Phienthrakuland Wanwisa Chevakulmongkol, "Handwritten Recognition on Pali Cards of Buddhadasa Indapanno", International Computer Science and Engineering Conference (ICSEC): ICSEC 2013

[10] Kavitha Subramani, Murugavalli Subramaniam, "Creation of original Tamil character dataset through segregation of ancient palm leaf manuscripts in medicine", 24 January 2020,DOI: 10.1111/exsy.12538

[11] Avadesh, Meduri, and Navneet Goyal. "Optical Character Recognition for Sanskrit Using Convolution Neural Networks" ,13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 447- 452. IEEE, 2018.

[12] Aqsa Rasheed, Nouman Ali,Amsa Shabbir, "Handwritten Urdu Characters and
Digits Recognition Using Transfer Learning and Augmentation With AlexNet", IEEE Access, Digital Object Identifier 10.1109/ACCESS.2022.3208959.

[13] T.S.Suganya , S.Murugavalli, "An Efficient Ancient Tamil Script Classification System using Gradient Boosted Tree Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ,ISSN: 2277-3878, Volume-8 Issue-3, September 2019

[14] Santoso, Rachmat & Suprapto, Yoyon & Yuniarno, Eko mulyanto, "Kawi Character Recognition on Copper Inscription Using YOLO Object Detection", 343-348. 10.1109/CENIM51130.2020.9297873.

[15] Deore, S.P., Pravin, A, "Devanagari Handwritten Character Recognition using fine-tuned Deep Convolutional Neural Network on trivial dataset", Sādhanā 45, 243 (2020).

[16] Suganya, T.S., Murugavalli, S, "A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition", Soft Comput 24, 10933–10941 (2020). https://doi.org/10.1007/s00500-019-04596-x.

[17] AditiMoudgil,SaravjeetSingh,VinayGautam,ShalliRani,SyedHassanShah, "Handwritten devanagari manuscript characters recognition using capsne", International Journal of Cognitive Computing in Engineering, Volume 4, June 2023, Pages 47-54
[18] R. Giridharan, E. K. Vellingiriraj and P. Balasubramanie, "Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning," 2016 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India, 2016, pp. 1-7, doi: 10.1109/ICRTIT.2016.7569600.
[19] T. Manigandan, V. Vidhya, V. Dhanalakshmi and B. Nirmala, "Tamil character recognition from ancient epigraphical inscription using OCR and NLP," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1008-1011, doi: 10.1109/ICECDS.2017.8389589.

