2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) | 979-8-3503-8564-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICV62344.2024.00022

Comparative Analysis of Outcomes of Tesseract OCR for Different Languages

Kartik Joshi
Assistant Professor
Faculty of Computer Applications and Information Technology
GLS University
Ahmedabad, Gujarat
ORCID ID: 0009-0009-4732-3804
kartikjoshi@glsuniversity.ac.in

Harshal Arolkar
Head
Faculty of Computer Applications and Information Technology
GLS University
Ahmedabad, Gujarat
ORCID ID: 0000-0003-0371-4466
harshal.arolkar@glsuniversity.ac.in

Abstract

Hindi and Gujarati are Devanagari scripts that have contributed to the culture of India in the form of literature and human interactions. Due to aging, the textually rich literature of these languages may not exist for future generations. To preserve them, the technological solution known as an optical character recognition engine is used. Tesseract OCR engine supports the conversion of Hindi and Gujarati language-based images to equivalent text outputs. Major challenge in generating better quality of text outputs is acquisition of high resolution and noise free images. This study has compared the working of the Tesseract OCR engine for the Gujarati and Hindi languages and determined how effective it is for real-time applications.

Keywords- Gujarati, Hindi, Optical Character Recognition (OCR), Tesseract

I. INTRODUCTION

Hindi, an Indo-Aryan language, is the most broadly communicated language in India and is likewise one of the authoritative dialects of the country. It is transcendently spoken in the northern areas of India. Hindi is derived from Sanskrit and has advanced over hundreds of years, integrating jargon and punctuation from different territorial dialects. The Devanagari script is utilized to compose Hindi. It is known for its rich abstract practice, with eminent artists and creators adding to Hindi writing. [1] Gujarati is also an Indo-Aryan language. It is predominantly spoken in the western Indian territory of Gujarat. It is additionally spoken by the Gujarati diaspora in different regions of the planet. Gujarati has its underlying foundations in Sanskrit and offers likenesses with other Indo-Aryan dialects. The language is written in the Gujarati script, which is derived from the Devanagari script. Gujarati writing has a long and regarded history, with popular scholars and artists adding to its rich legacy. [2]

Hindi as a language has 52 consonants and 13 vowels indicated with a Shiro Rekha, a line above characters, annotation used in many Devanagari scripts. Gujarati has 34 consonants and 12 vowels without the use of Shiro Rekha.

To save the ages-old literature, it needs to be converted into a digital copy for it to be re-printed and re-shared. The technical solution to do so can be done using the Optical Character Recognition engine. Tesseract is one of the world's most popular OCR engines available as open source. It started as a research project in Hewlett Packard labs situated in Bristol between 1985 to 1996. It was then delivered as open source in 2005 by Hewlett Packard and the College of Nevada, Las Vegas (UNLV). Tesseract advancement has been supported by Google since 2006 for multiple languages.

The Hindi language model in Tesseract OCR has 2745604 words, 412 punctuations, and 21 numbers. The Gujarati language model in Tesseract OCR has 98456 words, 403 punctuations, and 148 numbers in the relevant files

979-8-3503-8564-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ICICV62344.2024.00022

.

II. WORKING OF TESSERACT

Starting around 2005, there have been countless changes and improvements done in Tesseract OCR, however, in 2016 there was a colossal precision update brought into it utilizing the Long Short-Term Memory (LSTM) Model, at the cost of required computational power. Tesseract gives three types of LSTM-based models for most of the dialects, because of accuracy, speed, and combination of legacy Neural Networks (NN) and new LSTM mode [3] [4], it can be applied by setting Command Line argument 'oem'.

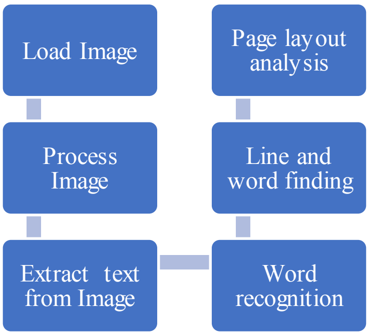
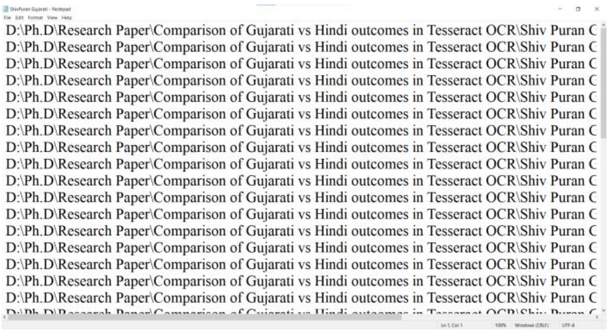


Figure 1: Workflow of Tesseract

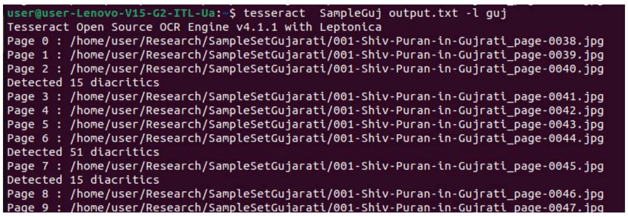
In Tesseract OCR 3.0x versions, text recognition is based on Long Short-Term Memory (LSTM). By simulating long-term dependencies in character sequences, the LSTM layer increased recognition accuracy.

The general function of LSTM is to handle text recognition sequentially while accounting for character dependencies and context within a word or line. Because LSTMs can capture long-range dependencies in sequences, they are helpful for text recognition in images where character order is important.



```
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
D:\Ph.D\Research Paper\Comparison of Gujarati vs Hindi outcomes in Tesseract OCR\Shiv Puran C
```

Figure 2: Text file of path and names of all Sample images for one language



```
user@user-Lenovo-V15-G2-ITL-Ua:~$ tesseract  SampleGuj output.txt -l guj
Tesseract Open Source OCR Engine v4.1.1 with Leptonica
Page 0 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0038.jpg
Page 1 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0039.jpg
Page 2 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0040.jpg
Detected 15 diacritics
Page 3 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0041.jpg
Page 4 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0042.jpg
Page 5 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0043.jpg
Page 6 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0044.jpg
Detected 51 diacritics
Page 7 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0045.jpg
Detected 15 diacritics
Page 8 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0046.jpg
Page 9 : /home/user/Research/SampleSetGujarati/001-Shiv-Puran-in-Gujrati_page-0047.jpg
```

Figure 3: Execution of Tesseract command

Run the command shown in Figure 3 to execute the operation.

Here, when we pass an image into the Tesseract engine as an input, the Tesseract OCR engine does page segmentation, layout analysis, line detection, and thresholding on an image, and passes that image data through the LSTM model to extract text from the image to generate text as output with help of supporting file compressed in ‘language’. trained data.[4] The list of files that are required for extracting text are listed here along with their purpose:

1. ‘Language’. lstm: This is an LSTM recognition model file generated by the training process, and this is the required file for the recognition process.

2. ‘Language’.lstm-unicharset: The file contains the Unicode character set that Tesseract recognizes, with properties. The same uni-charset must be used to train the LSTM.

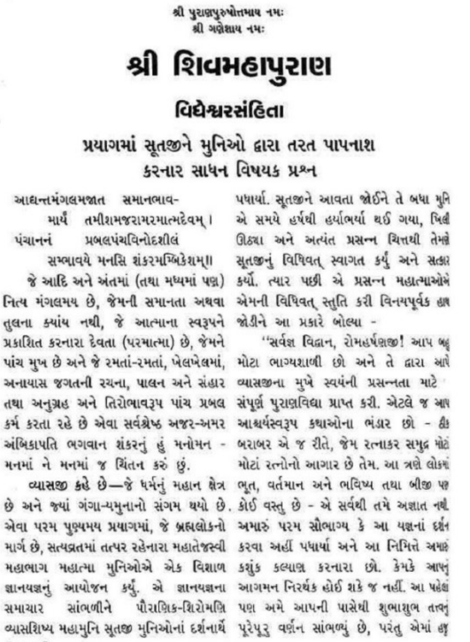
3. ‘Language’.lstm-recoder: It is a Uni-char-compressed (recorder), which maps the uni-charset further to the codes used by the neural network recognizer This is created as a part of the starter trained data by combine_lang_model.

4. ‘Language’.lstm-punc-dawg / ‘Language’.lstm-word-dawg / ‘Language’.lstm-number-dawg: Data Warehouse Advisory Group (DWAG) acts as supporting (optional) files and 'Language'.lstm-unicharset used to build the lstm-(punc | word | number)-dawgs files which are respectively punctuation, dictionary words and numbers or tokens, in place of "word" and digit space is added.

III. EXPERIMENTATION OF TESSERACT

As seen in Figure 3, it is possible can generate individual files for a set of respective images. For this experiment, the authors have selected the book “Shri Shiv Mahapuran”. This book is of huge historical significance in India and around the world. The book was originally written by Maharishi Ved Vyaasji around 10-11th century. This saga was published commercially for the first time in 1884 in Bombay (now Mumbai) then in 1896 in Calcutta (now Kolkata). The current edition was published first in 1906 and then revised in 1965 by Pandit Pustakalya, Kashi. The researchers have used these books to generate two different datasets of 900 images each. The books used as datasets for both languages, are at least 2 decades old and resemble an old type-writer-like font face instead of the new font face found in the current era of digital printing.

The two books were scanned using Canon PIXMA MegaTank GM4070 All-in-One to generate JPEG images for input from the Tesseract OCR engine. Figures 3(a) and 3(b) show sample images of the book for Gujarati and Hindi respectively.



શ્રી પુરાણપુરુષોત્તમાય નમઃ
શ્રી ગણેશાય નમઃ

શ્રી શિવમહાપુરાણ

વિદ્યેશ્વરસંહિતા

પ્રયાગમાં સૂતજીને મુનિઓ દ્વારા તરત પાપનાશ
કરનાર સાધન વિષયક પ્રશ્ન

આદ્યન્તમંગલમજાત સમાનભાવ-
માર્યં તમીશમજરામરમાત્મદેવમ્ ।
પંચાનનં પ્રબલપંચવિનોદશીલં
સમ્ભાવયે મનસિ શંકરમમ્બિકેશમ્॥

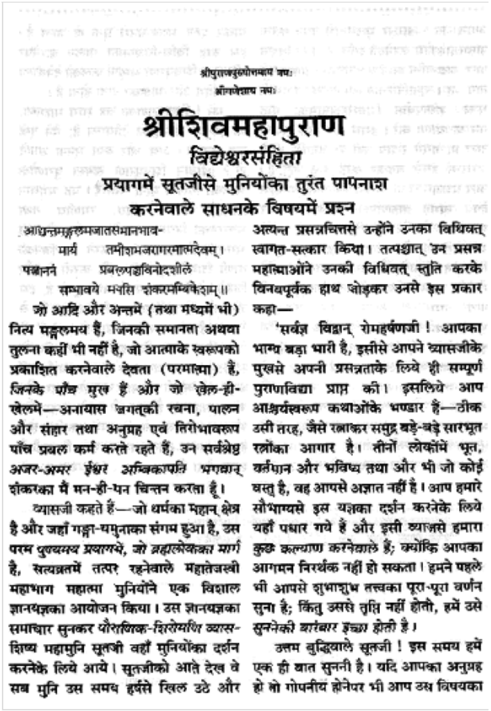
જે આદિ અને અંતમાં (તથા મધ્યમાં પણ) નિત્ય મંગલમય છે, જેમની સમાનતા અથવા તુલના ક્યાંય નથી, જે આત્માના સ્વરૂપને પ્રકાશિત કરનારા દેવતા (પરમાત્મા) છે, જેમને પાંચ મુખ છે અને જે રમતાં-રમતાં, ખેલખેલમાં, અનાયાસ જગતની રચના, પાલન અને સંહાર તથા અનુગ્રહ અને તિરોભાવરૂપ પાંચ પ્રબલ કર્મ કરતા રહે છે એવા સર્વશ્રેષ્ઠ અજર-અમર અંબિકાપતિ ભગવાન શંકરનું હું મનોમન - મનમાં ને મનમાં જ ચિંતન કરું છું.

વ્યાસજી કહે છે—જે ધર્મનું મહાન ક્ષેત્ર છે અને જ્યાં ગંગા-યમુનાનો સંગમ થયો છે. એવા પરમ પુણ્યમય પ્રયાગમાં, જે બ્રહ્મલોકનો માર્ગ છે, સત્યવ્રતમાં તત્પર રહેનારા મહાતેજસ્વી મહાભાગ મહાત્મા મુનિઓએ એક વિશાળ જ્ઞાનયજ્ઞનું આયોજન કર્યું. એ જ્ઞાનયજ્ઞના સમાચાર સાંભળીને પૌરાણિક-શિરોમણિ વ્યાસશિષ્ય મહામુનિ સૂતજી મુનિઓનાં દર્શનાર્થે પધાર્યા. સૂતજીને આવતા જોઈને તે બધા મુનિ એ સમયે હર્ષથી હર્યાભર્યા થઈ ગયા, ખિલી ઊઠ્યા અને અત્યંત પ્રસન્ન ચિત્તથી તેમણે સૂતજીનું વિધિવત્ સ્વાગત કર્યું અને સત્કાર કર્યો. ત્યાર પછી એ પ્રસન્ન મહાત્માઓએ એમની વિધિવત્ સ્તુતિ કરી વિનયપૂર્વક હાથ જોડીને આ પ્રકારે બોલ્યા -

‘‘સર્વજ્ઞ વિદ્વાન, રોમહર્ષણજી! આપ બહુ મોટા ભાગ્યશાળી છો અને તે દ્વારા આપે વ્યાસજીના મુખે સ્વયંની પ્રસન્નતા માટે સંપૂર્ણ પુરાણવિદ્યા પ્રાપ્ત કરી. એટલે જ આપ આશ્ચર્યસ્વરૂપ કથાઓના ભંડાર છો - ઠીક બરાબર એ જ રીતે, જેમ રત્નાકર સમુદ્ર મોટાં મોટાં રત્નોનો આગાર છે તેમ. આ ત્રણે લોકમાં ભૂત, વર્તમાન અને ભવિષ્ય તથા બીજી પણ કોઈ વસ્તુ છે - એ સર્વથી તમે અજ્ઞાત નથી. અમારું પરમ સૌભાગ્ય કે આ યજ્ઞનાં દર્શન કરવા અહીં પધાર્યા અને આ નિમિત્તે અમારું કશુંક કલ્યાણ કરનારા છો. કેમકે આપનું આગમન નિરર્થક હોઈ શકે જ નહીં. આ પહેલાં પણ અમે આપની પાસેથી શુભાશુભ તત્ત્વનું પૂરેપૂરું વર્ણન સાંભળ્યું છે, પરંતુ એમાં હજુ

[1286]

Figure 4(a): Shri Shiv Mahapuran in Gujarati



श्रीपुराणपुरुषोत्तमाय नमः
श्रीगणेशाय नमः

श्रीशिवमहापुराण

विद्येश्वरसंहिता

प्रयागमें सूतजीसे मुनियोंका तुरंत पापनाश
करनेवाले साधनके विषयमें प्रश्न

आद्यन्तमङ्गलमजातसमानभाव-
मार्यं तमीशमजरामरमात्मदेवम् ।
पञ्चाननं प्रबलपञ्चविनोदशीलं
सम्भावये मनसि शंकरमम्बिकेशम् ॥

जो आदि और अन्तमें (तथा मध्यमें भी) नित्य मङ्गलमय हैं, जिनकी समानता अथवा तुलना कहीं भी नहीं है, जो आत्माके स्वरूपको प्रकाशित करनेवाले देवता (परमात्मा) हैं, जिनके पाँच मुख हैं और जो खेल-ही-खेलमें—अनायास जगत्‌की रचना, पालन और संहार तथा अनुग्रह एवं तिरोभावरूप पाँच प्रबल कर्म करते रहते हैं, उन सर्वश्रेष्ठ अजर-अमर ईश्वर अम्बिकापति भगवान् शंकरका मैं मन-ही-मन चिन्तन करता हूँ।

व्यासजी कहते हैं—जो धर्मका महान् क्षेत्र है और जहाँ गङ्गा-यमुनाका संगम हुआ है, उस परम पुण्यमय प्रयागमें, जो ब्रह्मलोकका मार्ग है, सत्यव्रतमें तत्पर रहनेवाले महातेजस्वी महाभाग महात्मा मुनियोंने एक विशाल ज्ञानयज्ञका आयोजन किया। उस ज्ञानयज्ञका समाचार सुनकर पौराणिक-शिरोमणि व्यास-शिष्य महामुनि सूतजी वहाँ मुनियोंका दर्शन करनेके लिये आये। सूतजीको आते देख वे सब मुनि उस समय हर्षसे खिल उठे और अत्यन्त प्रसन्नचित्तसे उन्होंने उनका विधिवत् स्वागत-सत्कार किया। तत्पश्चात् उन प्रसन्न महात्माओंने उनकी विधिवत् स्तुति करके विनयपूर्वक हाथ जोड़कर उनसे इस प्रकार कहा—

‘सर्वज्ञ विद्वान् रोमहर्षणजी ! आपका भाग्य बड़ा भारी है, इसीसे आपने व्यासजीके मुखसे अपनी प्रसन्नताके लिये ही सम्पूर्ण पुराणविद्या प्राप्त की। इसलिये आप आश्चर्यस्वरूप कथाओंके भण्डार हैं—ठीक उसी तरह, जैसे रत्नाकर समुद्र बड़े-बड़े सारभूत रत्नोंका आगार है। तीनों लोकोंमें भूत, वर्तमान और भविष्य तथा और भी जो कोई वस्तु है, वह आपसे अज्ञात नहीं है। आप हमारे सौभाग्यसे इस यज्ञका दर्शन करनेके लिये यहाँ पधार गये हैं और इसी व्याजसे हमारा कुछ कल्याण करनेवाले हैं; क्योंकि आपका आगमन निरर्थक नहीं हो सकता। हमने पहले भी आपसे शुभाशुभ तत्त्वका पूरा-पूरा वर्णन सुना है; किंतु उससे तृप्ति नहीं होती, हमें उसे सुननेकी बारंबार इच्छा होती है।

उत्तम बुद्धिवाले सूतजी ! इस समय हमें एक ही बात सुननी है। यदि आपका अनुग्रह हो तो गोपनीय होनेपर भी आप उस विषयका

Figure 4(b): Shri Shiv Mahapuran in Hindi

The literature survey shows that Tesseract has been worked on with the existing Gujarati dataset to check its efficiency and working. The outcome of the study was that the Tesseract works well with the new-age printing type font face. The efficiency achieved was almost 93%.[6] [7] [8] Similar study has also been conducted for the existing Hindi dataset but for both, old as well as new type font faces and delivered efficiency is 90%. [9] [10]

The Tesseract can be trained for an entirely new language model thereby improving the overall efficiency of the respective language. There are other methods such as converting to black and white or shifting to the KNN method for better outcomes, but the pre-requisite of a complete dataset stays the same. [11]

The Tesseract OCR engine can generate a text file for each input image. The researchers have generated a single Excel file text file of all the output text files at once using a Python script.

As the Tesseract engine allows only to generate a text file of all the words identified in the given set of images, the researchers enhanced the experiment with the help of a Python script. The researchers were able to generate an Excel file with all the words and their occurrences in the entire dataset for each language, thus allowing additional analysis of each word and character alike. The generic steps used in the experiment are as follows:

1. Select the folder path containing all the text files generated by Tesseract OCR
2. Create an Excel file and call the functions sequentially to read, extract, remove duplicates, sort, and write the words to an Excel file to store all text from text files.
3. Read Text Files and Extract Words
4. Create a List of All Words
5. Remove Duplicates
6. Sort the List of Words based on their length.
7. Rewrite Sorted Words into Excel File

Repeat steps 1-5 for the next set of text files generated by Tesseract for Gujarati/Hindi language.

The generated Excel files were then compared manually letter by letter with the hardcopies of the book to understand the precision of the Tesseract OCR engine.

The authors have analyzed the efficiency of the language model based on individual characters irrespective of the vowels. If the given letter is included in a single, two-letter, three-letter, or more than three-letter word or with all possible vowels for the said letter, the identification outcome remains the same.

Here Table 1 shows the list of characters identified by Tesseract for Gujarati and Hindi languages whereas Table 2 shows the characters not identified by Tesseract for Gujarati and Hindi languages respectively.

Table 1: Character identification analysis for Gujarati and Hindi

Sr. No.	Character printed in Gujarati	Identified in Gujarati	Character printed in Hindi	Identified by Hindi
1	પ્ર	No	प्र	Yes
2	ધ્ય	No	ध्य	Yes
3	બ્ર	No	ब्र	Yes
4	મ્હ	No	म्ह	Yes
5	ત્ર	No	त्र	Yes
6	ધ્દ	No	ध्द	Yes
7	ગ્ર	No	ग्र	Yes
8	ભ્ર	No	भ्र	Yes
9	ષ	No	ष	Yes
10	દ્ર	No	द्र	Yes
11	વ્ર	No	व्र	Yes
12	સ્ત્ર	No	स्त्र	Yes
13	શ્ન	No	श्न	Yes
14	શ્વ	No	श्व	Yes

Table 2: Character identification analysis for Gujarati and Hindi

Sr. No.	Character printed in Gujarati	Identified in Gujarati	Character printed in Hindi	Identified in Hindi
1	ક	Yes	क	Yes
2	ખ	Yes	ख	Yes
3	ગ	Yes	ग	Yes
4	ઘ	Yes	घ	Yes
5	ઙ	Yes	ङ	Yes
6	ચ	Yes	च	Yes
7	છ	Yes	छ	Yes
8	જ	Yes	ज	Yes
9	ઝ	Yes	झ	Yes
10	ઞ	Yes	ञ	Yes
11	ટ	Yes	ट	Yes
12	ઠ	Yes	ठ	Yes
13	ડ	Yes	ड	Yes
14	ઢ	Yes	ढ	Yes
15	ણ	Yes	ण	Yes
16	ત	Yes	त	Yes
17	થ	Yes	थ	Yes
18	દ	Yes	द	Yes
19	ધ	Yes	ध	Yes
20	ન	Yes	न	Yes
21	પ	Yes	प	Yes
22	ફ	Yes	फ	Yes
23	બ	Yes	ब	Yes
24	ભ	Yes	भ	Yes
25	મ	Yes	म	Yes
26	ય	Yes	य	Yes
27	ર	Yes	र	Yes
28	લ	Yes	ल	Yes
29	ળ	Yes	व	Yes
30	વ	Yes	श	Yes
31	શ	Yes	ष	Yes
32	ષ	Yes	स	Yes
33	સ	Yes	ह	Yes
34	હ	Yes	क्ष	Yes
35	-	-	त्र	Yes
			ज्ञ	Yes

IV. OBSERVATION

From the above tables, it is observed that the existing Hindi dataset can identify all characters of the Hindi language along with vowels and joint characters. The current Gujarati dataset fails to identify 14 different types of characters along with vowels of the same character and joint characters.

1. The dataset fails to recognize Gujarati characters that have a certain curve in their design or vertices such as ઞ.

2. The dataset fails to recognize Gujarati characters which are made of two half characters such as પ્ર.

3. Due to a lack of training dataset and support for old-type font faces, the system fails to recognize these words. The same support has been given in the Hindi dataset for similar type font faces.

To improve the performance and accuracy of the Gujarati language, Tesseract OCR requires a trifecta of language-specific preprocessing, Tesseract parameter optimization, and training.

Custom language data can be used for training with Tesseract. Think about using your dataset to train the engine specifically for the Gujarati language. This includes setting up ground truth files, training the OCR engine, and producing training data. Although the procedure can be difficult and expert-only, it can greatly increase the accuracy of recognition.

To improve the quality of the input images, you can also use language-specific preprocessing methods. [12]

V. CONCLUSION

The authors have been able to perform a comparative analysis of two different outcomes generated for two different Indian languages respectively using Tesseract. Based on the outcome, it is possible to conclude that the efficiency of the Tesseract OCR engine is dataset-sensitive, the better it is trained, the better the outcome. It can also be concluded there is a need to create a new dataset for the Gujarati language which provides extensive support for multiple-type font faces, especially the older ones irrespective of size and color. Such a dataset will be proven helpful in the correct terms of saving old manuscripts written in Gujarati language.

REFERENCES

[1] “Hindi language – Britannica, “9 December 2023. [Online]. Available: https://www-britannica-com.translate.goog/topic/Hindi-language

[2] "Gujarati language - Wikipedia," 16 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Gujarati_language

[3] R. Smith, "An overview of the Tesseract OCR engine," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol. 2, pp. 629-633, 2007

[4] Sepp Hochreiter; Jürgen Schmidhuber *(1997)*. "Long short-term memory". Neural Computation. ***9*** *(8): 1735–1780*.

[5] "Tesseract OCR GitHub," 16 02 2017. [Online]. Available: https://github.com/tesseract-ocr.

[6] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open-source OCR tool tesseract: A case study," International Journal of Computer Applications, vol. 55, no. 10, 2012.

[7] M. Audichya, J. Saini, “A Study to Recognize Printed Gujarati Characters Using Tesseract OCR”, 1, 2 Computer Science, Gujarat Technological University

[8] C. Patel & A. Desai, “Gujarati Handwritten Character Recognition Using Hybrid Method Based on Binary Tree-Classifier And K-Nearest Neighbour”. International Journal of Engineering Research & Technology (IJERT), 2013. 2(6), 2337-2345.

[9] J. Kaur, V. Goyal & M. Kumar, “Improving the accuracy of tesseract OCR engine for machine printed Hindi documents”. AIP Conference Proceedings. 2022. 2455. 040007. 10.1063/5.0101164.

[10] N. Mishra, C. Patvardhan, V. Lakshimi and S. Singh, "Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition”. International Journal of Computer Applications, vol. 39, no. 6, pp. 19-23, 2012.

[11] C. Verstraeten, "How to train Tesseract 3.01 - Cédric Verstraeten," 16 02 2017. [Online]. Available: https://blog.cedric.ws/how-to-train-tesseract-301.

[12] T. Kundaikar, & J. Pawar,” Multi-font Devanagari Text Recognition Using LSTM Neural Networks”. 10.1007/978-981-15-0029-9_39. 2020