A Model for Identifying and Predicting deteriorated text in Manuscripts using *Vrutta* Pattern Matching

[1]Keertimalini A, [2]T. Venkata Padmaja, [3]Dr. B. Chandrasekaram

[1,2]P. G Students, [3]Associate professor,
Dept. of Computer Science,
National Sanskrit University, Tirupati 517 507.

Abstract: **Manuscripts are important cultural artifacts which provide significant insights into the history and culture of ancient India. Manuscripts form an invaluable part of India's documentary heritage. But due to years of time lapse, physical & chemical factors caused naturally or artificially by humans (foreign invaders) led to loss of valuable information. Manuscripts that may have undergone just a little amount of damage, which when identified through the remaining available text, can be used to extract the complete knowledge it contains. This paper presents one such model to predict and retrieve the deteriorated text from the Sanskrit manuscripts using metrical analysis and laghu-guru pattern matching techniques. This paper also extends the scope of prediction of deteriorated texts to metered and unmetered text and deals with the more efficient corpus searching algorithms using Regular expressions.**

Keywords: **Manuscript, Metrical analysis (Vruttas), Pattern matching, Chandas Shastra, Deterioration, Regular expressions, Sanskrit corpus.**

I. Introduction

1.1 Manuscripts - *Memory of the world*

National Mission for Manuscripts (NAMAMI), a project undertaking by the Ministry of Tourism and Culture defines a Manuscript as — "Handwritten composition on paper, bark, cloth, metal, palm leaf or any other material dating back at least seventy-five years that has significant scientific, historical or aesthetic value." Manuscripts are a treasure house of knowledge. It not just covers the traditional domains such as religion, culture & philosophy, but also contains innumerable scientific information on Medicine(like *Sushrutha and Charaka Samhita*), Astronomy and Astrology(*Jyothisa)*, Vedic Mathematics(Sulbasutras),Yoga(*Patanjali Yogasutras*), *Literature(Sahitya)*, History*(Purana-Itihasa), Language Grammar (Vyakaran) and Lexicons(Amarakosha):*India possesses one of the largest collections of manuscripts comprising more than five millions of manuscripts on different materials. They capture our thoughts, achievements, experience and lessons learnt from history; in other words, they constitute our 'memory.' UNESCO provided this recognition to the most valuable documentary heritage of the world - INDIA.Thus manuscripts form an important part of the Indian Knowledge system(IKS).

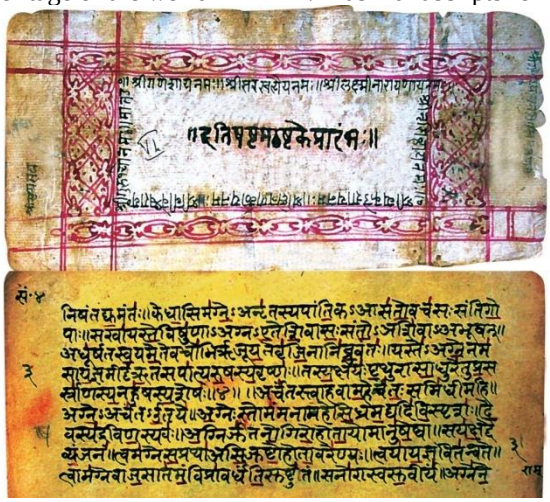

Fig. 1 Rigveda manuscript

Fig. 2 UNESCO Certificate

1.2 Deterioration of Manuscripts

Deterioration is a loss of structural capacity with time by the action of the external agents or material leaching (Saiz and Laiz, 2000). These aesthetically valuable resources, which are easily more than seventy five years old, are usually written on birch bark, palm leaf, handmade paper, cloth, leather etc. Found in various shapes and sizes, the deterioration of these collections can be in the form of wear and tear, shrinkage, cracks, warping, bio-infestation, discoloration, abrasion, holes,etc. The ravages of time, extreme

climatic conditions and biological agents often destroy priceless cultural property and records [1]. Here in this paper we deal with manuscripts which have lesser degree of deterioration(Point sized deterioration). The remaining text, when recognized, can help us retrieve the complete knowledge provided in the manuscript. This paper serves useful in preserving the archaic document and thus preserve the valuable knowledge it possesses.

Fig. 3 Deteriorated manuscript sample

1.3 Sanskrit Prosody - Chhandas Sastra & Vrutta ratnakara

Prosody or chhandas is the science of metres in Sanskrit literature. Metre is an important tool for poetical compositions. Sanskrit prosody was started with pingala's chhandas sastra, followed by Vrutta rathnakara of kedara and chandomanjari of Gangadasa. Chhandas sastra is one of the six vedangas or limbs of vedas.("Chandah padau tu vedasya"). These texts deal with metres which specify the name of the metre and its particular gana through verses i.e. via 1) Sutra form or 2) Karika form. Dandin, a proficient sanskrit poet remarks "Prosody is like a boat/ship for the reader who wants to cross the ocean of poetry."

Characteristics of Classical Metres:

The Sanskrit prosody is divided into two namely vedic and classical. Metres used in vedas are known as vedic, and metres used in sanskrit sahitya (*Purana Itihasa, Kavyas*) are known as classical.

a) **Pada (Quarter)** - Sanskrit shlokas or verses generally contain 4 lines, each of which is referred to as Pada. Depending upon the Vrutta chosen by the poet, each quarter is either characterized by matras or aksharas(Syllables).

b) **Akshara** - A sanskrit syllable which can be pronounced distinctly. Eg. रा, मः in रामः

c) **Maatra** - each syllabic instant or morae of an Akshara. Eg. र्+आ+म्+अः in the word रामः

d) **Laghu** - A syllable is short when the vowel is short. The vowels अ, इ, उ, ऋ, ऌ (a,i,u,r,lr) are short. It is denoted by the symbol ' **I** '

e) **Guru** - i) A syllable is long when the vowels are long. आ,ई,ऊ,ॠ,ॡ .

ii) A short vowel can become long when it is followed by anusvaaramअं(am) or visargamअः(aH).

iii) If a conjunct consonant - called samyuktaakShara in Sanskrit) follows a short vowel, then the short vowel becomes long.Except प्र, ब्र, क

iv) Any short vowel at the end of पाद.A guru syllable is denoted by the symbol '**U**'

f) **Vrutta** - Sanskrit metres that are classified according to the number of syllables occurring in each Paadam (quarter) of the verse.

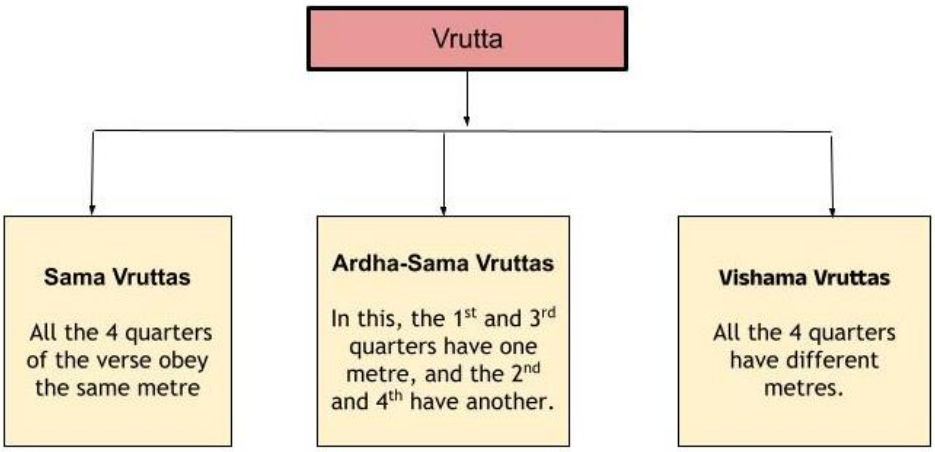

Fig. 4 Classification of Vruttas

g) **Gana** - A metrical feet is of combination of three syllables, which can either be guru or laghu. Thus 2^3 = 8combinations of ganas are possible.

आदिमध्यावसानेषुचरतायान्तिलाघवम्।
भजसागौरवंयान्तिमनौतुगुरुलाघवम्॥

This means, there is a short syllable at the start (aadi), middle (madhya) and end (avasaana) positions for the ganas य, र

1 Deterioration of Manuscripts and Indigenous Methods used for Preserving These Documentary Cultural Heritage Ravindra Goswami, Department of Botany, R.B.S. College, Agra.

and त respectively (the other 2 syllables being long). Similarly, for भ,ज, and स there is a long syllable in the same positions (the other 2 syllables being short). Ma and na are fully long and short respectively.

" यमाताराजभानसलगं " - is usually the acronym used to refer to these ganas.

यगण (ya)	रगण (ra)	तगण (ta)	भगण (bha)	जगण (ja)	सगण (sa)	मगण (ma)	नगण (na)
IUU	UIU	UUI	UII	IUI	IIU	UUU	III

Table 1 The 8 Ganas and their patterns

1.4 Regular Expressions

A regular expression (also known as regex) is a sequence of characters that defines a search pattern used for matching or manipulating strings of text. It can be used to search for specific patterns or sequences of characters within a larger text, and can be used in a variety of programming languages and applications.Regular expressions can include special characters and symbols, which can be used to specify certain types of characters or patterns. Here we have used a basic regular expression '([IU]*)' to find and return the missing laghu-guru pattern of the deteriorated zone.Using regular expressions, the efficiency of searching and pattern matching of laghu-guru mappings are increased and so is the accuracy of output as well.

1.5 Pattern matching

According to DeepAI Pattern matching is defined as **"**An algorithmic task that finds pre-determined patterns among sequences of raw data or processed tokens."Using regular expressions, the given data is checked and then matched with a process-of-elimination approach, like backtracking. This task makes exact matches from the existing *Vrutta* database and returns the pattern of the missing text.

1.6 Sanskrit Corpus

Sanskrit corpus data is a collection of Sanskrit text which contains text extracted from the epics like Ramayana, Mahabharata, Bhagavad Gita and Vedic texts like Vedas. In this paper the corpus data is used to extract the words which are predictable in the deteriorated zones. Some of the corpus data from SanskNET was collected and saved as files. When we find the pattern for the deteriorated zone, say P(A), the corresponding word is returned from this corpus via enhanced corpus searching methods.

II. Literature Survey

The restoration of damaged manuscripts has been a topic of interest for many researchers. Different works have been carried out related to the restoration and analysis of damaged manuscripts, as well as a system developed for Sanskrit verse metrical analysis.This paper has been mainly inspired from the research paper of Dr.B.Chandrasekharam, Associate Professor, National Sanskrit University titled " Automatic digital rebuilding of text from deteriorated zones of manuscripts containing metered text.The author proposes a model via which it may be possible to minimize the loss of text, for the deteriorated zone of a manuscript which contains metered text. This is done by taking advantage of Chhanda sastra. It proposes a model/tool for rebuilding text from deteriorated/corrupted zones of manuscripts containing metered text.

Apart from this, the Jñānasaṅgrahaḥ system platform developed at IIT Kanpur, hosts an application for Sanskrit verse metrical analysis. This application allows users to obtain the metrical analysis of Sanskrit verses. Techniques for restoring damaged manuscripts via image processing proposed by Sai Siddharth Kota et al and the digital restoration of erased and damaged manuscripts, published by Keith T et al, are some prominent works in this arena. Another such work, authored by Diwakar Mishra, is titled "Strategies for Metrical Analysis of Sanskrit Text" and provides details about the metrical analysis of Sanskrit texts.

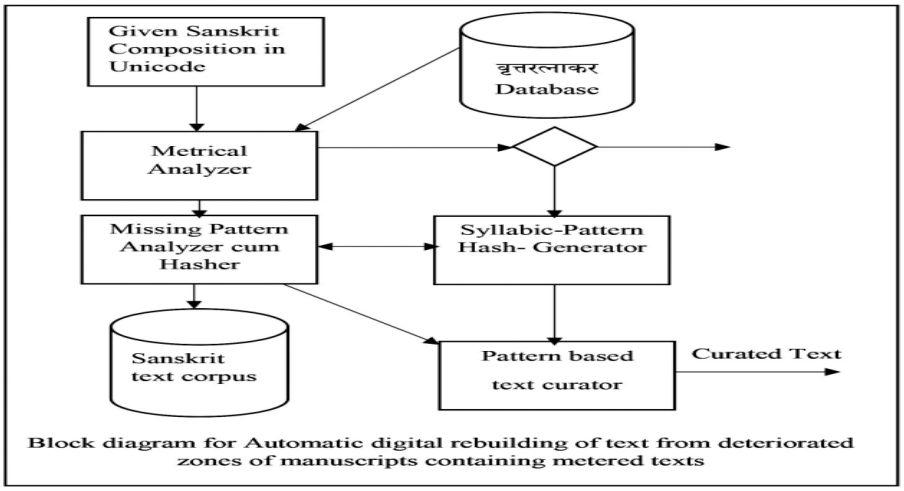

Fig. 5 Block diagram of the previous Model

III. Methodology

Manuscripts which have been deteriorated need to be pre-processed to extract the information from it. For identifying the available content and marking the deteriorated zones, we first digitize the manuscript and upload it in the web interface that has been created.

Step 1 - Digitization of manuscript and image enhancement

In this web application there is an image input box where the user will input a corrupted manuscript image. The noise is removed from the raw image using Guassian bandpass filtering as suggested in the paper "Digital Enhancement of Indian Manuscripts."[2] The enhanced image is now of the readable form to the users.

Step 2 - Keying in of Unicode Sanskrit Text and marking missing zones

Below the image input box there is a text input box. There is a technique called Optical Character Recognition (OCR) Technology which will convert the image into editable text. But here the user is going to input the text manually in the text box and mark the missing or deteriorated parts in capitalized alphabets (A,B,C etc.) The text will be in Devanagari Unicode text format.

Process involved in the above 2 steps

1: Input - Sanskrit shloka which contains some deterioration.

माताÓमत्पितारामचÓ
स्वामीÓमत्सखारामचन्द्रः।

2: Combine both lines into a single line.

माताÓमत्पितारामचÓस्वामीÓमत्सखारामचन्द्रः।

3: Label the deteriorated zones using alphabets.

मातAमत्पितारामचBस्वामीCमत्सखारामचन्द्रः।

Step 3 - Mapping of laghu guru:

Using the laghu-guru assigning process that has been explained above, the laghu-guru patterns of the given input text is found. The source code has been designed in such a way that the words will be split into Aksharas which will then be patternised into binary laghu-guru pattern. The output of the inserted shloka will be represented as:

Input :

मातAमत्पितारामचBस्वामीCमत्सखारामचन्द्रः।

Output:

U U A U I U U I U B U U C U U I U U I U U

```
def display_laghu_guru(txt):
    char_split=list(split_clusters(txt))#UUU
    laghu_list=
    guru_list2=
    laghu_str=
    guru_str=
    deergamatras=
    laghumatras=

    final_list=''
    pattern = re.compile(r'^[A-Za-z]$')
    for i in char_split:
        #print(i+'--'+str(len(i)))
        if pattern.match(i):  ## added 3 line to check whether given given char is aplbet or add  to deteriorated string
            final_list=final_list+i
            continue  #end of the code
        if len(i)==1:
            if i in laghu_str:
                final_list=final_list+'I'
            elif i in guru_str:
                final_list=final_list+'I'

        elif  len(i)==2 :
            j=i[-1]
            if j in laghu_str:
                final_list=final_list+'I'
            elif j in guru_str:
                final_list=final_list+'U'
            elif j in laghumatras:
                final_list=final_list+'I'
            elif j in deergamatras:
                final_list=final_list+'U'
        elif  len(i)>=3 and char_split [-1]!=i:
            final_list=final_list[:-1]
            final_list=final_list+'UU'

        #print(final_list)
    final_str=final_list
```

[2]Digital Enhancement of indian Manuscript, Yashodhar Charitra , Sai siddarth Kota, Raja Massand, Abhinaya agrawal and Preety Singh

Fig. 6 Source code for laghu-guru mapping

Step 4 : Detection of P(A) via pattern matching

To identify the exact I,U pattern denoted as P(A), for deteriorated text A, firstly a database containing a set of Vrutta patterns (*Sama, Ardha-sama, Vishama*) is created. Through this database, the P(A) is found by comparing the P(I) with the entries of the data set. The chandas of the shloka can also be found. The following steps will assist in understanding the process in detail:-

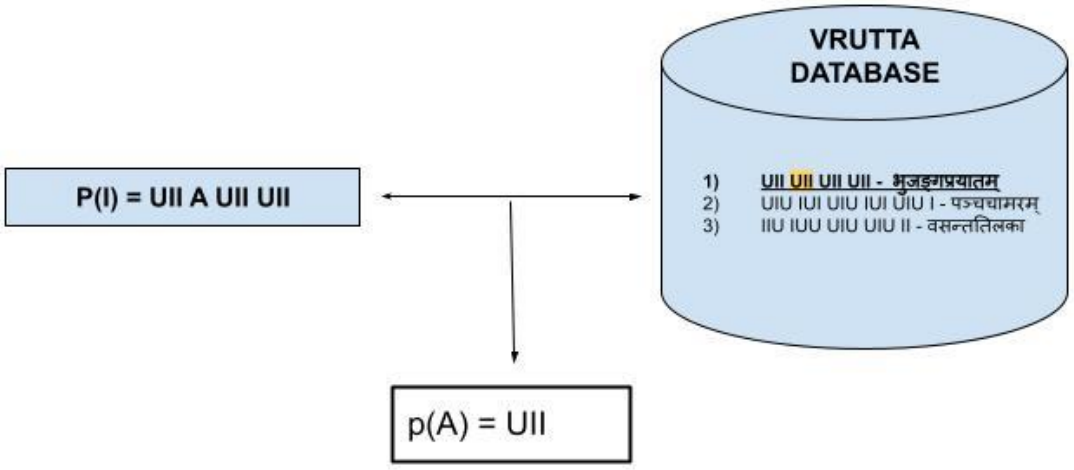

Fig. 7 Pattern matching process

1) Creating a database of containing Vruttas and its corresponding pattern.

	S.NO	NAME OF THE VRUTTA	GANA	NO. OF SYLLABLES IN EACH PADA	NO. OF PADAS	PATTERN	PATTERN
सम वृत्ताः	1	शालिनी	मततगग	11	4	III IIU IIU II III IIU IIU II III IIU IIU II III IIU IIU II	III IIU IIU II III IIU IIU II III IIU IIU II III IIU IIU II
	2	भुजङ्गप्रयातम्	यययय	12	4	UII UII UII UII UII UII UII UII UII UII UII UII UII UII	UII UII UII UII UII UII UII UII UII UII UII UII UII UII
	3	पञ्चचामरम्	जरजरजग	13	4	UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I	UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I UIU IUI UIU IUI UIU I
	4	इन्द्रवज्र	ततजगग	11	4	IIU IIU UIU II IIU IIU UIU II IIU IIU UIU II IIU IIU UIU II	IIU IIU UIU II IIU IIU UIU II IIU IIU UIU II IIU IIU UIU II
	5	उपेन्द्रवज्रः	जतजगग	11	4	UIU IIU UIU II UIU IIU UIU II UIU IIU UIU II UIU IIU UIU II	UIU IIU UIU II UIU IIU UIU II UIU IIU UIU II UIU IIU UIU II

Fig. 8 Sama Vrutta Database

	S.NO	NAME OF THE VRUTTA	NO. OF SYLLABLES IN 1st and 3rd PADA	NO. OF SYLLABLES IN 2nd and 4th PADA	GANA OF 1ST AND 3RD PADA	GANA OF 2ND AND 4TH PADA	NO. OF PADAS	PATTERN	PATTERN
अर्ध - सम वृत्ताः	1	हरिणीप्लुता	11	12	सस्स+ IU	नभ्भर	4	IIU IIU IIU IU III UII UII UIU IIU IIU IIU IU III UII UII UIU	IIU IIU IIU IU III UII UII UIU IIU IIU IIU IU III UII UII UIU
	2	औपच्छन्दसिक पुष्पिताग्रा	12	13	ननरय	नजजर +U	4	III III UIU IUU III IUI IUI UIU U III III UIU IUU III IUI IUI UIU U	III III UIU IUU III IUI IUI UIU U III III UIU IUU III IUI IUI UIU U
	3	विद्योगिनी	10	11	ससजग	सभर+IU	4	I I IUI U II UII UIU IU II UII UIU IU	I I IUI U II UII UIU IU II I IUI U II UII UIU IU
	4	अपरवक्त्र	11	12	ननरलग	नजजर	4	III III UIU IU III IUI IUI UIU III III UIU IU III IUI IUI UIU	III III UIU IU III IUI IUI UIU III III UIU IU III IUI IUI UIU
	5	उपचित्र	8	10	सससलग	भभभगग	4	I I I I IU UII UII UII UU I I I I IU UII UII UII UU	I I I I IU UII UII UII UU I I I I IU UII UII UII UU
	6	वेगवती	7	11	सससग	भभभगग	4	I I I I U UII UII UII UU I I I I U UII UII UII UU	I I I I U UII UII UII UU I I I I U UII UII UII UU
	7	वसन्तमालिका	9	11	ससजगग	सभरय	4	I I I IUI UU I UII UIU IUU I I I IUI U I UII UIU IUU	I I I IUI UU I UII UIU IUU I I I IUI U I UII UIU IUU

Fig. 9 Ardha-Sama Vrutta Database

2) By having the above data set we will compare P(I) with the predicted match pattern. If any of the patterns is similar to the existing patterns then it will identify the missing pattern.

Input pattern:

U U A U I U U I U B U U C U U I U U I U U

Existing pattern which is similar:

U U U U U I U U I U U U U U U I U U I U U

Output: Identified pattern for deteriorated zone: A= UU, B= U, C= UU

Step 5: Prediction of missing text of the identified pattern via a Corpus:

Prediction of the missing text A of the deteriorated manuscript is identified using corpus searching algorithms in the given Sanskrit Corpus.

Here is an algorithm which is used to speed up searching in the Sanskrit corpus using regular expression:

Syntax:

```
import re
corpus = open("sanskrit_corpus.txt", "r").read()
regex = re.compile(pattern)
matches = re.findall(regex, corpus)
 for match in matches:
        print(match)
```

1. As shown above, a particular text can be chosen which needs to be given as input.
2. Specify the particular regular expression that you want to search in the corpus. I
3. Here few Sanskrit Unicode data texts has been inserted into a filed, which acts as the corpus. This makes easy access to the corpus so that code can search it efficiently.
5. Once the searching is done, an array of similar type of words which match the deteriorated pattern P(A) is returned as output. In some cases, it takes the first word as reference and gives a similar type of output.

Expression which is used in the source code:

```
def find_chars(input_str):
   regex_pattern = r'[^IU]'
   # Use the re.findall() method to find all the matches other than IU
   matches = re.findall(regex_pattern, input_str)
    return matches
```

FINAL OUTPUT:

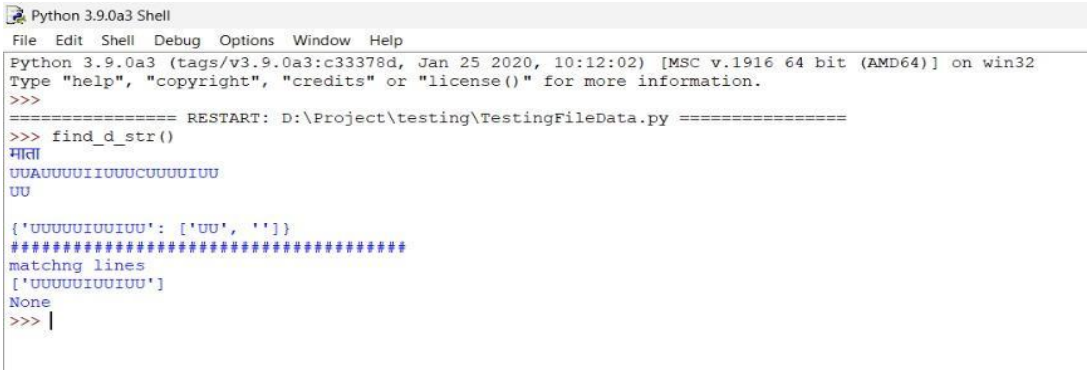

Fig. 10 Output displaying laghu-guru mappings of input

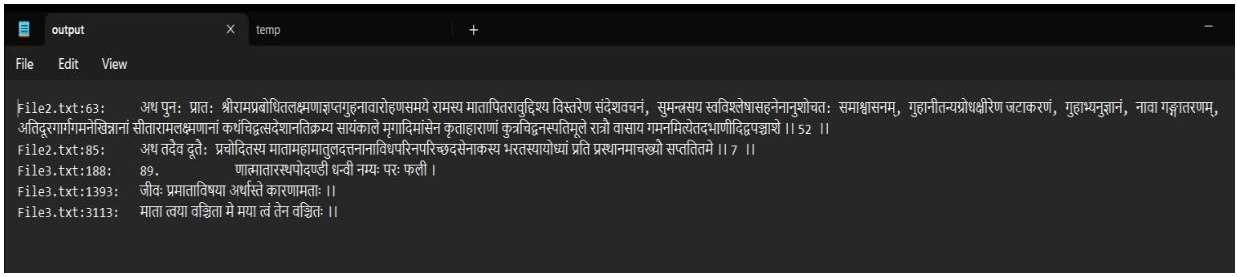

Fig. 11 Output displaying the predicted missing text

IV. Conclusion

In the field of manuscript studies “The model for identifying and predicting deteriorated text in manuscripts using Vrutta pattern” matching will become an efficient tool for researchers and scholars. Through this Vrutta pattern matching becomes easier to identify the text in the deteriorated zone which helps in preserving the valuable historical documents.This method has the potential to greatly enhance the understanding of our archaic texts and knowledge treasure found in the manuscripts. The scope for improvement of this model is sky high. The continued use and refinement of this model is likely to yield even more significant results in the future.

REFERENCES:

[1]. Deterioration of Manuscripts and Indigenous Methods used for Preserving These Documentary Cultural Heritage Ravindra Goswami, Department of Botany, R.B.S. College, Agra.
[2]. Digital Enhancement of Indian Manuscript, Yashodhar Charitra ,
[3]. ASHWINI S. DEO (2007). The metrical organization of Classical Sanskrit verse. Journalof Linguistics, 43, pp 63114 doi:10.1017/S0022226706004452

[4]. Natarajan Meghanathan et al.:WiMONe, NCS, SPM CSEIT-2014 pp.199-207, 2014 doi:10.5121/csit.2014.412216
[5]. Internet resource:https://www.researchgate.net/publications/265644163
[6]. Roger L., Easton.Jr, Keith.T.Knox, "Digital Restoration of Erased and Damaged Manuscripts", Proceedings of the 39th Annual Convention of the Association of Jewish Libraries (Brooklyn, NY-June 20-23,2004)
[7]. https://sanskrit.iitk.ac.in/jnanasangraha/chanda/verse
[8]. Pingala, Chandas-Sastra, Kavyamala series no.91 (3rd Edition) Bombay, 1938.
[9]. RAMA N. AND Meenaksh Lakshmanan, A Computational Algorithm for Classification of Verse, IJCSI International Journal Computer Science Isssues, Vol.7, Issue 2, No.1 March 2010, ISSN:1694-078