

TEXT BASED SANSKRIT LANGUAGE IDENTIFICATION

THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE

SUBMITTED BY

VIPIN TONGAR (2K20/CSE/25)

UNDER THE SUPERVISION OF

PROF. RAJNI JINDAL



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

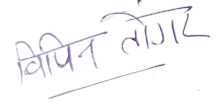
Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Vipin Tongar**, Roll No. **2K20/CSE/25** student of M.Tech (Computer Science), hereby declare that the Thesis Dissertation titled “**Text Based Sanskrit Language Identification**” which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is bona fide and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.



Place: Delhi

Student: Vipin Tongar

Date: 27/05/2022


DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Thesis Dissertation titled “**Text Based Sanskrit Language Identification**” which is submitted by Vipin Tongar, Roll No. 2K20/CSE/25, to Department of Computer Science, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 27/05/2022


Prof. Rajni Jindal

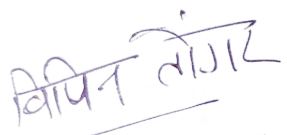
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDEMENT

I wish to express my sincerest gratitude to Dr Rajni Jindal for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its relevance. She was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been completed.

Place: Delhi

Date: 27/05/2022


Student: Vipin Tongar

Abstract

Text Based Identification of a language is the process of automatically detecting a certain language based on the text given in an article or document. Language identification is an established domain of research that has received considerable attention in the past. Language identification is a crucial initial step in various other works of Natural language processing, language translation, performing language specific AI models etc.

It is somewhat easier to differentiate languages which do not belong to same language family or not having same script because the characteristics features seldom overlap each other and due to different script, the symbols used to express the text are different. In this paper, we will devise method to identify Sanskrit language among various languages having same Devanagari script.

Contents

Candidate Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Content	v
List of Figures	vi
List of Abbreviations	vii
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 LITERATURE REVIEW	4
CHAPTER 3 METHODOLOGY	7
3.1 Comparing Symbols	7
3.2 Comparing Words	8
CHAPTER 4 IMPLEMENTATION	10
CHAPTER 5 RESULT AND EVALUATION	12
CHAPTER 6 CONCLUSION	14
REFERENCES	15
LIST OF PUBLICATIONS	17

List of Figures

Fig 3.1: Devanagari Unicode chart	7
Fig 3.2: Steps to identify Sanskrit Language	9
Fig 5.3: Bar chart showing accuracy percentage according to no. of unique words in list	13

List of Abbreviations

LID- Language Identification

NLP- Natural Language Processing

CED- Character Encoding Detection

SVM- Support Vector Machine

1. INTRODUCTION

Language Identification is the process of determining the natural language in which the document or text written in. It is a subfield of Natural Language Processing. It is one of the most extensively researched domain of NLP. LID is the key part in text processing pipelines as the text processing techniques already assumes that the language of the text is known before applying these methods. Language identification has fascinated computational linguists for many decades. There are various methods which evolved with advanced techniques to solve such problem. Several statistical approaches to language identification using different techniques to classify the data are present. One method is to compare the compressibility of the text with the compressibility of texts in a set of known languages. This approach is known as mutual information based distance measure. The same technique can also be used to empirically construct family trees of languages which closely correspond to the trees constructed using historical methods. Mutual information based distance measure is essentially equivalent to more conventional model-based methods and is not generally considered to be either novel or better than simpler techniques.

Another technique, which is more advanced and still used today is the n-gram model technique devised by Trenkle and Cavnar in 1994 to create a model which first trains from the training text of the each of the languages and then is used to identify languages. These models can be based on characters or encoded byte. In character based the language identification and character encoding detection(CED) are integrated. Then, for any piece of text needing to be identified, a similar model is made, and that model is compared to each stored language model. The most likely language is the one with the model that is most similar to the model from the text needing to be identified. This approach can be problematic when the input text is in a language for which there is no model. In

that case, the method may return another, "most similar" language as its result. Also problematic for any approach are pieces of input text that are composed of several languages, as is common on the Web.

With growing social media presence on internet and interaction of multilingual people on these platforms, the identification of language based on the text typed has become an area of immense importance. There are now tons of data available to users in hundreds of major languages spoken by millions of people worldwide. The automatic processing of these texts, for any purpose that requires NLP, such as indexing, querying requires the identification of the primary language. LID further helps the machine to translate these texts by identification of the language used and for various other aspects related to it.

The paper deals with identification of Sanskrit language. Sanskrit Language is not only a language with literary tradition spanning thousands of years but also a predecessor for many languages spoken by people in today's world. Nearly all Indian languages directly or indirectly derive many linguistic aspects from this ancient and beautiful language.

Nearly all religious texts of Hinduism, Jainism and Buddhism are written in it. Immense knowledge and experience of our past generations is concealed in the form of books in this language. With growing popularity of Sanskrit among present generation and its reach in the internet, the need of identifying the texts of it has become more significant. It not only helps in various other fields of Natural language processing but most importantly it would make it easier to automatically translate Sanskrit texts once they are identified.

The script used in writing Sanskrit texts is Devanagari. It is one of most used writing system or script in the world as various Indo-European languages such as Hindi, Marathi, Nepali uses it. There are different types of methods applied in

language identification such as Naïve based classification, neural network, n-gram method, markov model, support vector machine(SVM) etc.

The method that we are going to use is elimination method. Each language has some unique characteristics which include certain alphabets, words that are not used in any other language. If we have to differentiate a particular language from other languages of same writing system or script, we have to check whether these characteristics lies in the text. If they lie, then we can say the text does not belongs to that particular language because it is containing unique features of some other language. In this way we can pinpoint whether text belongs to the language we are identifying or not. Our objective is to create Sanskrit language identification system using this method and also comparing these in terms of probability of accurately predicting the language of the text.

2. LITERATURE REVIEW

A lot of research has been done in this field and significant advancements have been made in this field over the past decade. Various methods have been evolved and practiced in the precise detection of language such as n-gram method, markov model, naïve-based classifier, neural networks etc. We are going to discuss recent developments and studies carried out in this field. Tools and algorithms related to pattern recognition field were also used for language detection but outcomes have shown that simpler and classification methods are difficult to be outperformed by these methods. Hence, n-gram is still widely used for this purpose.

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n - 1)$ order Markov model. n-gram models are now widely used in probability, communication theory, computational linguistics, computational biology, and data compression. Two benefits of n-gram models and algorithms that use them are simplicity and scalability – with larger n, a model can store more context with a well-understood space–time tradeoff, enabling small experiments to scale up efficiently.

Some methods can easily take into account hundreds of languages and can predict with high accuracy that the text given is of that particular language from the list of languages it can identify or sort, while some are specifically designed or manipulated to work across a particular language or a language family. Various researches have been published earlier on both of these types. The work of Gerrit Botha on South African languages is commendable. He used n-gram method for classification and presented a language identification system for these South African languages. The identification of some of the most frequently and distinctive sequences of words for any language was the important part of the system.

Many researchers discussed and compared various important methods according to accuracy of result in detection of language. Tommi Vatanen compared the two significantly used methods i.e. ranking method by Trenkle with Naïve Bayes classifier applied with n-gram model. Aditya Bhargava showed that n-gram model is better in detection or identification of language than all other models in language modeling. Some research are also done on Indian languages in this regard.

One of the great bottlenecks of language identification systems is to distinguish between closely related languages. Similar languages like Bulgarian and Macedonian or Indonesian and Malay present significant lexical and structural overlap, making it challenging for systems to discriminate between them.

These works have helped us to know the difficulties encountered while formulating language identification and techniques of applying various methods related to this field.

For better understanding of the elimination method which we are going to use, it was important to have basic knowledge of Devanagari script languages so that we can identify unique characteristics of each one of them which subsequently will help in identifying possibility that the language is Sanskrit or not. The languages considered among languages of Devanagari script were Hindi, Sanskrit, Marathi and Nepali as these 4 languages are mostly used while Bhojpuri, Konkani and others using Devanagari script have lesser reach on internet as well as lesser speakers.

Hindi is fourth most spoken language in the world. It is one of the official languages of India with significant presence in majority of states in India. It also has wide reach on internet as compared to other Indian languages. Sanskrit is classical language of India with gigantic literary works spanning thousands of years. Many Indian languages are derived from it. It acts as mother languages of

many Indian languages. Marathi is the state language of Maharashtra and have millions of speakers worldwide. Nepali is the official language of Nepal. It also has significant numbers of speakers as well as tons of digital information available in Nepali on internet.

In this work, we attempt to create model for identification of Sanskrit language among various Devanagari script languages. We will create an elimination model which identify whether language is indeed Sanskrit or not based on presence of unique characteristics of other Devanagari languages.

3. METHODOLOGY

The method is elimination method in which unique characters and words of languages are listed out. The symbols and words in the given text are compared whether they contain unique feature of any language other than Sanskrit. If this happens, then the possibility of text in Sanskrit language is eliminated. First, we compare alphabets and if no result is shown then we proceed to comparing words in the text.

3.1 Comparing alphabets

Each letters or symbols belonging to Devanagari script are assigned unique Unicode. Unicode is an Information Technology standard used for representing and encoding texts in all the writing systems in the world. The Devanagari Unicode block ranges from U+900 to U+97F and contains all required symbols of the script as shown in figure below.

Devanagari ^[1]																	
Official Unicode Consortium code chart (PDF)																	
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
U+090x	ं	ँ	ं	ः	अ	अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ँ	ऐ	ए	
U+091x	ऐ	ऑ	ओ	ओ	औ	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	
U+092x	ठ	ड	ढ	ण	त	थ	द	ध	न	न	प	फ	ब	भ	म	य	
U+093x	र	र	ल	ळ	ळ	व	श	ष	स	ह	ं	ी	ः	ऽ	ा	ि	
U+094x	ी	ु	ू	ृ	ृ	ँ	े	े	ै	ॉ	ो	ो	ौ	्	ि	ौ	
U+095x	ँ	ं	ं	ँ	ँ	ँ	ं	ं	ं	क	ख	ग	ज	ड	ढ	फ	य
U+096x	ऋ	ॠ	ॡ	ॢ	।	॥	०	१	२	३	४	५	६	७	८	९	
U+097x	०	१	अँ	अं	आँ	आं	अु	अू	रु	ज़	ष	ग	ज	?	ड	ब	

Fig 3.1: Devanagari Unicode chart

There are various symbols which are used in other languages but do not find place in Sanskrit. For example, 'nuqta' (क़, ख़, ग़, ज़, ड़, ढ़, फ़, य़, ऩ, ऱ) are used in Hindi

frequently due to Persian words influence, but there are no such symbols in Sanskrit. Similarly, ॉ ॐ ॐ are used in Hindi and other languages to represent many words of English or European languages but are not used in Sanskrit.

फ़ is used in Marathi similar to sound of letter ल but is not seen in any other language of Devanagari script.

‘:’ symbol which is called visarga is used in Sanskrit language placed after noun words but is rarely seen in any other language we are considering. There are very few instances where visarga is used in Hindi and also as punctuation symbol. We can consider that if visarga symbol comes at end of any word then the text can be in Sanskrit. But we can’t solely rely on visarga to state that language is Sanskrit. So, we have to proceed on comparing words even if visarga is present.

So, the text whose language is to identified contains such letters which do not find place in Sanskrit, then it can easily be inferred that the text is not in Sanskrit language and we can reach not have to compare words.

3.2 Comparing words

As Sanskrit has considerable influence over many Indian languages particularly Devanagari script languages, there are lot of similarities between these languages. But, there are still many features which are unique in a language. This also applies to these 4 languages. Hindi, Marathi and Nepali share many characteristics with Sanskrit but there is still significant amount of words which are confined to a particular language only.

If in a given text which is to identified whether or not it belongs to Sanskrit languages, it is important to analyze whether this text have any word that uniquely belongs to some other language. If this is the case, then surely the text is not of Sanskrit. To proceed in developing the system, it is required to compile a list of unique words in all these 3 languages. Each word in the text is compared with the

words in the list. If found, then the text is ruled out as being the text of not Sanskrit language.

In Hindi there are various words which are unique. Few of them are: है, जो, और, हैं, गए, गई, हो, था, उनके, इन, यह, वाले, चुके, इसका, होता, वह, हुई, अपने, भी

In Marathi few unique words are: आहे, आणि, गेले, त्याला, तुमचा, सुद्धा, गेली, सदर, येथे, आले, दोन, त्यांची, झाले, नऊ, वाटप, आहेत

Similarly, in Nepali some of the unique words are: छ, छन्, र, यसको, गर्दछ, हुन्, भएको, पनि, तर, मान्दछ, एउटा, ठूलो, थियो

A list is created which contains considerable amount of such unique words of these languages. The possibility of the text to be identified correctly became high with more such words.

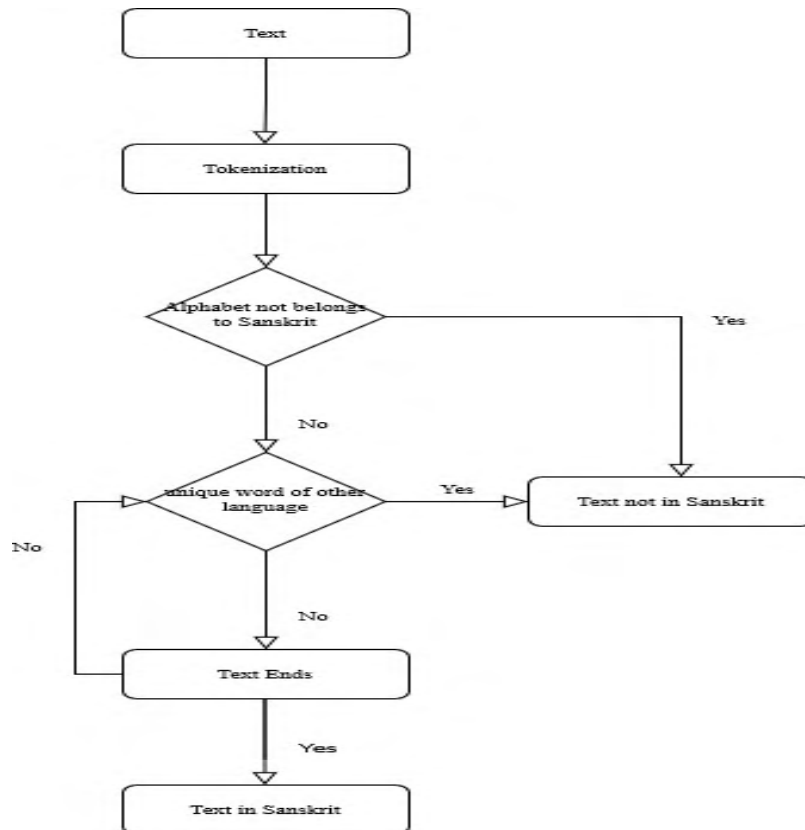


Fig 3.2: Steps to identify Sanskrit Language

4. IMPLEMENTATION

The implementation of the method that is designed to identify the Sanskrit language is done using the python programming language. The part where comparison of symbols is concerned is not being able to implement due to abugida (segmental writing system in which consonant-vowel sequences are written as units) nature of Devanagari script. The tokenize method for non-English language does not lead to desired result. But the method to compare words in the text with the list of unique words of 3 languages i.e. Hindi, Marathi, Nepali worked as per the expectations.

Code

```
import nltk
```

```
text= "अपने दोस्तों को आमंत्रित करें"
```

```
word_list=["है","जो","और","हैं","गए","गई","हो","था","उनके","इन","यह","वाले","  
चुके","इसका","होता","वह","हुई","अपने","भी","आहे","आणि","गेले","त्याला","तुमचा  
","सुद्धा","गेली","सदर","येथे","आले","दोन","त्यांची","झाले","नऊ","वाटप","आहेत"  
","छन्","र","यसको","गर्दछ","हुन्","भएको","पनि","तर","मान्दछ","एउटा","ठूलो","  
थियो"]
```

```
spl=text.split();
```

```
l=0
```

```
for i in spl:
```

```
    if i in word_list:
```

```
print("The text is not in sanskrit");  
l=1  
break  
if l==1  
break  
  
if(l==0)  
print("Text is in Sanskrit");
```

5. RESULT AND EVALUATION

The system developed using the given method which involves simple approach of elimination works well. Most of the phrases and sentences were correctly decided.

Many texts which contained alphabets alien to the Sanskrit were simply identified by using alphabet comparison approach. If in the text there were no alien symbols, the next step of comparing words of the text and the list containing unique words of other 3 languages to identify whether unique words of these languages are present or not was initiated. Initially the list of words of Nepali, Marathi and Hindi languages were taken in less quantity. It was increased subsequently in definite proportion and change in outcomes were noted.

Initially the probability of correctly guessing text as being Sanskrit or not was somewhat less. With increase of words in the list, the probability also increased but increment was comparatively large initially. The slope of accuracy with respect to list size was steep first and then eventually it become gentle.

Many a times the result was shown in lesser time. It was when alphabets not belonging to Sanskrit are found and system running ended there. The accuracy initially was about 64% which increased till 92% by increasing list size eventually. The chart is shown below which graphically shows how accuracy changes with list size.

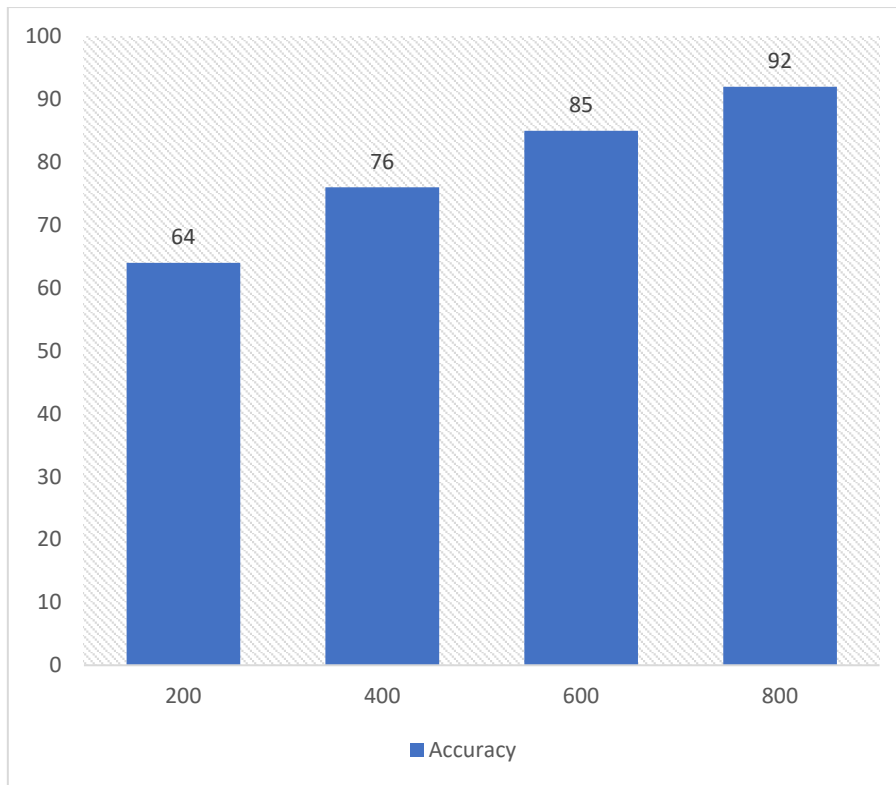


Fig 5.3: Bar chart showing accuracy percentage according to no. of unique words in list

6. CONCLUSION

The study shows that Language identification can be performed easily with simpler and effective method. Sanskrit serves as vocabulary source for many Indian languages, the amount of loan words is considerably more in languages sharing Devanagari script. Hindi itself borrows more than 70% words from Sanskrit and similarly others also borrows significant amount of words. This acts as barrier in our method as the amount of unique words that remains are quite less in number as compared to entire word collection of a language. Still the method yield result with decent accuracy. The accuracy also varies based on nature of texts used for identification and list size. The alphabets which do not find place in Sanskrit also enhances the accuracy in slight amount.

There is still much to progress in this regard. The tokenization of text into symbols is uphill task in abugida languages such as Sanskrit. But if the tokenization is dealt, it will make it easier to detect the language rapidly and in less time. There is still more ways to improve the method and enhance the accuracy of its application.

REFERENCES

- [1] Indhuja K, Indu M, Sreejith C and P. C. Reghu Raj, "Text Based Language Identification System for Indian Languages Following Devanagiri Script," *International Journal of Engineering Research & Technology*, vol. 03, no. 04, 2014.
- [2] Sreejith, C., M. Indu, and PC Reghu Raj. "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts," *In Fourth International Conference on Computing, Communications and Networking Technologies, IEEE*, pp. 1-4, 2013.
- [3] Jauhiainen, Tommi, Krister Lindén, and Heidi Jauhiainen. "Language model adaptation for language and dialect identification of text." *Natural Language Engineering*, vol. 25, no. 5, pp. 561-583, 2019.
- [4] Chew, Yew Choong, Yoshiki Mikami, and Robin Lee Nagano. "Language identification of web pages based on improved n-gram algorithm." *International Journal of Computer Science*, vol. 8, no. 3, pp. 47-58, 2011.
- [5] Ramu Reddy, V., Sudhamay Maity, and K. Sreenivasa Rao. "Identification of Indian languages using multi-level spectral and prosodic features." *International Journal of Speech Technology*, vol. 16, no. 4, pp. 489-511, 2013.
- [6] Kumar, V. Ravi, Hari Krishna Vydana, and Anil Kumar Vuppala. "Significance of gmm-ubm based modelling for indian language identification." *Procedia Computer Science*, vol. 54, pp. 231-236, 2015.
- [7] G Radha Krishna, R Krishnan, V K Mittal, "An Automated System for Regional Nativity Identification of Indian speakers from English Speech", *IEEE 16th India Council International Conference (INDICON)*, pp.1-4, 2019.
- [8] Shivesh Ranjan, Chengzhu Yu, Chunlei Zhang, Finnian Kelly, John H. L. Hansen, "Language recognition using deep neural networks with very limited training data", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5830-5834, 2016
- [9] Niraj Kr. Singh, Anoop Singh Poonia, "Analysis of prosody based automatic LID systems", *Communication, Control and Intelligent Systems (CCIS)*, pp.152-156, 2015.

- [10] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, Pedro Moreno, "Automatic language identification using deep neural networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5337-5341, 2014.

LIST OF PUBLICATION

Vipin Tongar and Rajni Jindal. “Text Based Sanskrit Language Identification”.
In: Proceedings of Industrial Electronics & Electrical Engineers Forum. April
2022. pp. 5-9.

Abstract

Text Based Identification of a language is the process of automatically detecting a certain language based on the text given in an article or document. Language identification is an established domain of research that has received considerable attention in the past. Language identification is a crucial initial step in various other works of Natural language processing, language translation, performing language specific AI models etc.

It is somewhat easier to differentiate languages which do not belong to same language family or not having same script because the characteristics features seldom overlap each other and due to different script, the symbols used to express the text are different. In this paper, we will devise method to identify Sanskrit language among various languages having same Devanagari script.

PAPER NAME

thesisVipin.pdf

WORD COUNT

2959 Words

CHARACTER COUNT

15285 Characters

PAGE COUNT

25 Pages

FILE SIZE

320.8KB

SUBMISSION DATE

May 19, 2022 9:20 PM GMT+5:30

REPORT DATE

May 19, 2022 9:21 PM GMT+5:30**● 15% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 13% Internet database
- 0% Publications database
- Crossref Posted Content database
- 14% Submitted Works database

● Excluded from Similarity Report

- Crossref database
- Bibliographic material
- Cited material
- Small Matches (Less than 10 words)