Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry

Milind Kumar Audichya
Computer Science
Gujarat Technological University
Ahmedabad, Gujarat, India
milindaudichya@gmail.com

Jatinderkumar R. Saini
Computer Science
Gujarat Technological University
Ahmedabad, Gujarat, India
saini_expert@yahoo.com

***Abstract*—Metadata generation for the poems based on the unified rules is very complex from the viewpoint of computational linguistics. This is more tedious when it comes to Hindi poems. Prosody or 'Chhand' as it is called in the Hindi language consists of several sets of rules and which are used while construction of a Hindi poem. Currently, no such metadata generator or technique is in existence which can generate the metadata of Hindi poetry based on the prosody or any other Hindi grammatical rule. In this research paper, we are trying to wrap up all the difficulties with respect to the metadata generator based on 'Chhands', Types of 'Chhands', Parts of 'Chhands', Classification of 'Chhands'. Along with this paper majorly focuses on the unified-rule based technique for the generation of metadata based on the different set of rules of prosody. On 3026 different UTF-8 based inputs of poems, part of poems, stories, etc., We were able to achieve 98.09% accuracy with the implementation of this unified-rule based technique of metadata generation, rest 1.91% failure was mostly due to input data related issue.**

Keywords— Chhand, Hindi Poetry, Metadata Generator, Prosody, Metre, Computational linguistic

I. INTRODUCTION

"Hindi" which is written in Devanagari script, is an official language of the Republic of India. [1] There is a rich heritage of poetry in Hindi literature which is consist of so many decades. For an instance of this in the light of the fact that the following Concluding Doha of Hanuman Chalisa is just a core level of the types of Hindi Poetry and which was penned in 16th Century by poet Goswami Tulsidas.[2]

"पवनतनय संकट हरन, मंगल मूरति रूप।
राम लखन सीता सहित, हृदय बसहु सुर भूप॥"

The foundation parts or components of the majority of Hindi poetry are as follows:

- Figure of speech (*अलंकार*)
- Sentiment (*रस*)
- Verse or Metre (*छंद*)

These components generate especial magic in Hindi poetry which is quite enough to touch the heart of everyone no matter whoever the person who is just reading or reciting or listening to the poem. All of these above three components are having their own set of rule and regulations. Each component itself is having several types and further subtypes and so on. To streamline the process of the research work we choose to work with the bottom-up approach on the Verse or Metre which is also known as 'Chhand' (*छंद*) or 'Chhands' in Hindi, we used the term 'Chhand' and 'Chhands' for explaining prosody based rules in this whole paper. There is an important place of prosody in Hindi Poetry writing. Prosody is coming from the long back of the ancient era. Due to the lack of proper or systematic documentation of rules, regulations, and unavailability of creations examples, prosody is losing its charm and about to extinct if not taken care properly today.

II. LITERATURE REVIEW

Computational linguistics related research work can be found of different languages such as Arabic, Bangle, Chinese, English, Malay, Punjabi, Sanskrit and Spanish. [3] [4] Many research works related to specifically in Sanskrit based Computational Linguistics can be explored. [3] J. Kaur and JR. Saini introduced "PuPoCI", A Punjabi poetries classifier with the use of weighting and linguistic features. [4] Rule-based modeling of the research work is practiced by a group of researchers for the clustering of the word based upon rules for extracting metadata of the document. [5]

The uses of automatic metadata play a vital role in the various thing that can be understood very well by a joint research work of several researchers on building a structured syllabus repository. [6] In legal areas also metadata is useful to get the content description from legal information which is researched by semantic scholar A. Gangemi, M. Sagri, and D. Tiscornia. [7] Computational linguistic-based metadata building research (CLiMB) for automatically identifying and categorizing and some more aspect related to image metadata was another group of researchers. [8]

If we see into poetry segment specifically J. Kaur and JR. Saini tested 10 different Machine Learning algorithms for the classification of Punjabi Poetry. [9] As per the literature

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

review still there is a need for computational linguistic-based metadata generation related research works in Hindi Poetry, so we are looking forward to accomplishing this by contributing efforts in this research work.

III. The Problem

This section is consisting of the problems or issues which we are going to deal with in this research work. With reference to the Introduction, now we are already aware that we are working with prosody only we will focus on the problems regarding the same.

The major difficulties which we face when it comes to prosody in Hindi poetry are as follows:

A. The absence of previous research work
B. Contradictory Information
C. Unavailability of knowledge
D. Nested Complexity Levels
E. Datasets and Implementation Strategy

A. The absence of previous research work

One of the very first problems of this research work which makes this research work more challenging is there are no prior research works specifically related Chhands as on date is either not available or not done.

B. Contradictory Information

There is no standard source of information related to 'Chhands'. There may be much reason behind that but few are as below:

- No specific in-depth rules are easily available due to lack of arranged documentation.
- Due to lack of arranged documentation, the poets didn't follow all the rules unanimously which are defined long back in the ancient era due to that one may find the different rules for the same type of 'Chhand' at different places.

C. Unavailability of knowledge

There are very few sources or experts available from where the knowledge of the prosody can be obtained. Some efforts are made to add 'Chhands' in the curriculum of Hindi Grammar but unfortunately, that's just an abstract part of 'Chhands' or some limited number of 'Chhands' are included.

There is no provision or any specialization in this area of study of prosody, which can be a part of the curriculum of Higher Study.

D. Nested Complexity Levels

The reason why people do not choose to work within this field of research is consist of the nested complexity levels. As we already discussed that the information and knowledge are not available so easily and if it's available somewhere than also it's contradictory.

Apart from that if 'Chhands' are studied in depth there are types, types of types (subtypes), types of subtypes and so on. For each type or subtype formation rules and regulation gets changes so it's quite challenging to handle everything together.

E. Datasets and Implementation Strategy

Till now we are aware that there are so many challenges to deal with this research area. So along with those challenges, one more and very important fact is that due to the absence of any previous research works or related research works and all the issue mentioned above there were no ready dataset or implementation strategies are there by which a researcher can follow a specific path.

These things together make this research unique and tough to deal with. We are giving our level best to make this research work and all other works related to this research area profound and robust so this can open a new wing in the field of the computational linguistics research branch.

We highly believe in focusing on solutions more rather than focusing on problems. In the upcoming very sections of this research paper, we are going to describe the overview of prosody and much more in-depth findings related to 'Chhands' which we found during carrying out this research work.

IV. The Objective

The main motive or objective of this research work is to properly structure and standardize the scattered knowledge regarding prosody which is available in either deficient manner or contradictory at different sources of information.

Along with that with the point of view of computational linguistics, the research work is also focused to mold, a set of standardized rules, for the automatic generation of the metadata based upon these standard unified rules of prosody.

V. The Overview of prosody in Hindi Poetry

To understand the methodology of the implementation work of this research work we need to know about the prosody first. We are trying to include every penny information we obtained from various trustworthy sources like some precious ancient books [10] [11], online portals [12], blogs [13]–[17] some handwritten notes, etc. [18] And after the validation through some experts suggestion on contents and understanding of the facts and rules related to various aspects of prosody. Let's begin the journey to explore prosody.

According to the ancient books, we came to know that prosody is coming long back from Vedas and Puranas which are an Integral part of Hinduism ideology which is considered as oldest religion in the world.

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

- We found the presence of the above-mentioned fact as well when we explored more about this. Prosody is one of the six '***Vedangs'*** which are auxiliary disciplines associated with the study of '***Vedas'*** in Vedic culture which was developed in ancient times.
- Along with that, we found the presence of information related to 'Chhands' in '***Agni Purana'*** which is one of the eighteen 'Puranas'. The core information related to prosody and other arts of writing can be seen in 'Agni Purana' from Chapter 328 to 347. All this information is in Sanskrit which is an old Indo-Aryan language.
- 'Chhandsharstra' which provides knowledge of creating prosody which is discussed and referenced in 'Vedas' and 'Puranas' as well was created by a saint and great old mathematician named Pingal, who is also known as 'Pingalacharya' and 'Sheshavtar'.
- The poetical composition is called Verse or Metre or 'Chhand', which is not poetical is called prose.

A. Types of 'Chhands'

There are two main types of 'Chhands' which are as follows:

a. 'Vedic Chhands'
b. 'Laukik Chhands'

a) 'Vedic Chhands'

Vedic Chhands are coming from Vedas or we can say that uses of the Vedic Chhands can be seen in the Vedic Sanskrit Mantras. Some of the Vedic Chhands are 'Atyashthi' *(अत्यष्टि)*, 'Atijagti' *(अतिजगती)*, 'Atishakkari' *(अतिशक्वरी)*, 'Anushtup' *(अनुष्टुप)*, 'Ashti' *(अष्टि)*, 'Ushnik' *(उष्णिक्)*, 'Ekpada Virat' *(एकपदा विराट)*, 'Gayatri' *(गायत्री)*, 'Jagati' *(जगती)*, 'Trishthup' *(त्रिष्टुप)*, 'Dwipada Virat' *(द्विपदा विराट)*, 'Dhruti' *(धृति)*, 'Pankti' *(पंक्ति)*, 'Pragath' *(प्रगाथ)*, 'Prastar Pankti' *(प्रस्तार पंक्ति)*, 'Bruhati' *(बृहती)*, 'Mahabruhati' *(महाबृहती)*, 'Virat' *(विराट)*, 'Shakkari' *(शक्वरी)*.[19]

b) 'Laukik Chhands'

'Laukik Chhands' are the 'Chhands' which are created by the people or poets and are not part of Vedas. These are not specifically dependent on any rules strictly but rhythm must be maintained. Majority of modern newcomers and some old poets as well creates their poems without the knowledge of prosody rules which are defined by 'Chhandsharstra'. The reason behind that is there is no specific book or guide is available for the same. And in case if some books are available than also, they are too complex with least clarity, incomplete, contradictory and confusing. This is also a strong reason for carrying out this research work and our main or primary focus is on 'Laukik' Hindi 'Chhands' only which are widely used in Hindi Poetry.

B. Part of 'Chhands'

There are six major parts of a 'Chhand' which are as follow:

a) 'Charan' or 'Pad' (चरण या पाद)
b) 'Varna' and 'Matra' (वर्ण और मात्रा)
c) 'Yati' (यति)
d) 'Gati' (गति)
e) 'Tuk' (तुक)
f) 'Gana' (गण)

Let us know about each quickly.

a) 'Charan' or 'Pad' (चरण या पाद)

A 'Chhand' usually have four 'Charans'. 'Charan' is one-fourth part of 'Chhand'. Each 'Charan' or 'Pad' have a fixed number of 'Varns' or 'Matras' (Diacritics & Character Count). Some 'Chhands' may have more than four 'Charans' / 'Pads'.

There are two types of 'Charan':

1. 'Sam-Charan' *(समचरण)*: 2nd and 4th 'Charans' are called 'Sam-Charan'.
2. 'Visham-Charan' (विषमचरण):1st and 3rd 'Charans' are called 'Visham-Charan'.

Here is an example to understand this in a better way:

रहिमन धागा प्रेम का, मत तोड़ो छिटकाय |
टूटे से फिर ना जुड़े, जुड़े गाँठ पड़ जाय ||

Here are 'Charans' / 'Pads' of this above 'Doha'.

1st: *रहिमन धागा प्रेम का,*
2nd: *मत तोड़ो छिटकाय |*
3rd: *टूटे से फिर ना जुड़े,*
4th: *जुड़े गाँठ पड़ जाय ||*

Here 1st and 3rd are 'Visham-Charan', and 2nd and 4th are 'Sam-Charan'.

b) 'Varna' and 'Matra' (वर्ण और मात्रा)

The time duration of pronouncing any 'Varna' is known as 'Matra'. 'Varna' is also having two types. Each verse is composed of a combination of 'Varnas' or characters. There are two main types of characters:

1. 'Laghu' *(लघु)*: The one whose symbol is (l) and quantity is 1. The character in which it takes a little time to pronounce is called a 'Laghu Varna'.
   The following Characters are considered as 'Laghu Varna':
   अ, इ (ि), उ (ु)
2. 'Guru' *(गुरु)*: The 'Guru' whose symbol is (s) and quantity is 2. The character in which it takes more time to pronounce is called a 'Guru Varna'.
   The following Characters are considered as 'Guru Varna':
   आ (ा), ई (ी), ऊ (ू), ए (े), ऐ (ै), ओ (ो), औ (ौ), अं (ं), अः (ः)
3. 'Plut' *(प्लुत)*: There is a third character too which is called 'Plut', which takes more time to pronounce

than a Guru, its purpose is in musical composition. There are three quantities of 'Plut'.

Apart from these, there are several more aspects, rules, regulation, and some exception too where 'Laghu' is considered as 'Guru' and 'Guru' became 'Laghu'. We will discuss this in the upcoming section.

c) 'Yati' (यति)

While reading a 'Chhand', where the reader takes a pause or a small stop, is called 'Yati' *(यति).* Some symbols are fixed for 'Yati' which are as follow:

" , ", " | ", " | | ", " ? ", " ! "

d) 'Gati' (गति)

When reading the 'Chhand', the reader experiences a rhythm or flow, it is called 'Gati' *(गति)* or speed.

e) 'Tuk' (तुक)

The frequency of the characters at the end of the 'Charan' is called 'Tuk' *(तुक).*

f) 'Gana' (गण)

'Varnik Chhand' *(वर्णिक छन्द)* calculation is based on the particular sequence of 'Gana'. There is a 'Gana' of three 'Varna'. There are eight 'Ganas'.

The eight 'Ganas' are:

1. *यगण (।ऽऽ)*, 2. *मगण(ऽऽऽ)*,
3. *तगण (ऽऽ।)*, 4. *रगण(ऽ।ऽ)*,
5. *जगण (।ऽ।)*, 6. *भगण(ऽ।।)*,
7. *नगण (।।।)*, 8. *सगण(।।ऽ)*

There is a 'Ganasutra' *("यमाताराजभानसलगा")* to know a specific 'Gana' take one letter from the first eight characters 'Ganasutra' and add 'Gana' in the end. To know specific 'Gana' just take three 'Matras' including that character after applying 'Guru-Laghu' Matras in 'Ganasutra'.

Example:

। ऽ ऽ ऽ । ऽ ।। । ऽ

य मा ता रा ज भा न स ल गा

य + गण = यगण (।ऽऽ),

ज + गण = जगण (।ऽ।)

This is how we can know about any of the eight 'Ganas'. The upcoming table shows how to we can use this 'Ganasutra' along with the example word of each 'Gana'.

TABLE I. 'GANAS' WITH EXAMPLES

Key	*'Gana'*	*Symbol*	*Example*
(य)	*यगण*	*।ऽऽ*	*कमाना*
(मा)	*मगण*	*ऽऽऽ*	*आजादी*
(ता)	*तगण*	*ऽऽ।*	*आजाद*
(रा)	*रगण*	*ऽ।ऽ*	*कामना*
(ज)	*जगण*	*।ऽ।*	*नवीन*
(भा)	*भगण*	*ऽ।।*	*रावण*
(न)	*नगण*	*।।।*	*रमन*
(स)	*सगण*	*।।ऽ*	*रचना*
(य)	*यगण*	*।ऽऽ*	*कमाना*

C. Classification of 'Chhands'

There are many kinds of classification can be found in different sources but for this research work, we will be following the classification of Hindi 'Chhands' in three classes.

First two classes are having a rich set of Chhands and the rest all the free hand creation comes in the last third classification. The classification followed by this research work is decided after the in-depth exploration of Hindi Poetry. The classification of Hindi Chhands is as following:

a) 'Matrik Chhands' (मात्रिक छंद)
b) 'Varnik Vrutt / Chhands' (वर्णिक वृत्त /छंद)
c) 'Mukt / Muktak Chhands' (मुक्त / मुक्तक छंद)

These three classes are further divided into the subclasses and so on. We will know about each more in detail.

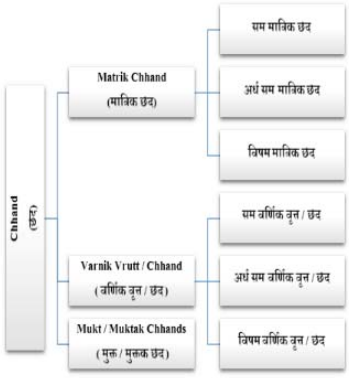

Figure 1. 'Chhand' Classification

The above Fig.1 shows an in-depth classification of the 'Chhand'. In which we can see that there are three levels of major classes for both 'Matrik Chhand' and 'Varnik Vrutt / Chhand' which are 'Sam', 'Ardh-Sam' and 'Visham' 'Chhands'.

a) 'Matrik Chhands' (मात्रिक छंद)

The verses in which the number of quantities *(मात्रा)* are fixed in all 4 stanzas are called 'Matrik' verses. The sequence of 'Varna' ('Laghu'/ 'Guru') doesn't matter in this type of verses.

Example:

1 2 1 2 2 1 1 1,1 2 1 1 2 1 1 2 2[11,13]

।ऽ ।ऽऽ ।।।,।ऽ ।। ऽ।। ऽऽ

तुम्हें वरेगी विजय, अरे यह निश्चय जानो।

2 1 1 2 1 1 2 1, 2 1 2 2 2 2 2[11,13]

ऽ।। ऽ ।। ऽ।, ऽ।ऽ ऽऽ ऽऽ

भारत के दिन लौट, आयगे मेरी मानो।।

The above 'Chhand' is a 'Rola' which an example of 'Matrik Chhand'. Here if you see matra count of each 'charan' is calculated based on the quantity of 'Laghu-Guru'. After that, each 'charan's' Quantities sum is written in the box [11,13].

The rule for 'Rola' is each even 'Charan' must have 13 Quantities or 'Matras', and each odd must have 11 Quantities or 'Matras'. Along with that, there must be a 'Guru' at the end of even 'charans'. So, this is just one example of 'Matrik Chhands'. The rule differs for every other 'Matrik Chhand' as per its creation regulations.

b) 'Varnik Vrutt / Chhands' (वर्णिक वृत्त /छंद)

The verses which are based on the calculation of 'Varnas', and the sequence of 'Varnas' ('Laghu'-'Guru') is fixed in all 4 stanzas are called 'Varnik Vrutt / Chhands'. The 'Ganas', 'Ganasutra' and 'Ganas' calculation plays a vital role in these kinds of 'Chhands' creation.

Example:

।ऽ । ऽ ऽ।।ऽ ।ऽ। ऽ,।ऽ। ऽ ऽ ।। ऽ। ऽ। ऽ
1 2 1 2 2 1 1 2 1 2 1 2, 1 2 1 2 2 1 1 2 1 2 1 2,
वसन्त ने सौरभ ने पराग ने, प्रदान की थी अति कान्त भाव से ।
।ऽ । ऽ ऽ।।ऽ ।ऽ। ऽ,।ऽ। ऽ ऽ ।। ऽ। ऽ। ऽ
1 2 1 2 2 1 1 2 1 2 1 2, 1 2 1 2 2 1 1 2 1 2 1 2,
वसुन्धरा को पिक को मिलिन्द को,मनोज्ञता मादकता मदान्दता ॥

This is 'Vanshasth Varnik Chhand' in which if we see each of its 'charan' is having a combination of some 'Ganas'. The rule of 'Vanshasth' is: there must be a sequence of 'Ganas' as 'JaGana', 'TaGana', 'Jagana', and 'RaGana'. If you see in 'Ganasutra' Table 1. You can observe that 'Jagana' *(।ऽ।)*, 'TaGana' *(ऽऽ।)* and 'RaGana' *(ऽ।ऽ)*.

Now if we arrange these sequence as per the rule of 'Vanshasth' it will become like: *(।ऽ।ऽऽ।।ऽ।ऽ।ऽ)* which should be followed in each 'charan' of 'Chhand'. So, this is just one example of 'Varnik Chhands'. The rule differs for every other 'Varnik Chhand' as same as 'Matrik' as per its creation regulations.

c) 'Mukt / Muktak Chhands' (मुक्त / मुक्तक छंद)

The verses which are not based on any of the above rules and which doesn't classify under 'Matrik' or 'Varnik' came into this classification. In this kind of verses calculation of 'Varnas' and Number of Quantities of 'Matras' depends on the creator's creation. They may or may not derive rules from existing 'Varnik' or 'Matrik Chhands'.

On a lighter note, we can say that all the freestyle or unrestricted creation of 'Chhands' are classified into this category. "Juhi Ki Kali" *(जूही की कली)* by a poet Suryakant Tripathi 'Nirala' is considered as the best creation in this 'Mukt' or 'Muktak Chhands' classification category.

VI. Methodology

For automatic metadata generation from the Hindi Poetries based on the unified rules of prosody composing we followed the bottom-up approach as the rules and criteria differ the most at the last level or core level.

The most important thing of this approach is the detection of 'Chhands' takes place very first and after that detection we put it in its appropriate type and subtypes as it becomes systematic to manage as types and subtypes usually are known to us already so we just need to map that detected 'Chhand' into its predefined or pre-stored type or subtypes.

The significant advantage of this approach is with small variations we can incorporate changes of rules of detection of 'Chhands', the addition of new 'Chhands' with ease of the implementation and get desired quality results. The upcoming paragraph explains about the automatic metadata generation technique's core methodology in depth.

The metadata generator's major work is to accept a particular part of the poem such as a couple of lines or a set of lines in UTF-8 Standard format for Devanagari [20] Script for the Hindi language. These number of lines may differ from 'Chhand' to 'Chhand' as per the construction rules of prosody which usually vary according to different 'Chhand's' core types. Ultimately the metadata generator expects the inputted data to be a 'Chhand' so it can complete its processing and provide appropriate metadata related to that 'Chhand' including all the relevant information in a well-organized manner.

We are going to discuss how the inputted data gets processed in this paragraph. After taking input very first thing, which is performed on the data is it's trimming and cleaning the junk parts from the data which can be considered as preprocessing of the entered data. Preprocessed data is now ready for further course of action. In the very next stage after preprocessing the preprocessed data now get separated line by line very carefully. The default delimiter for line separation is considered is newline character '\n' which can be changed too if required by making some nominal changes. After the separation of the data line by line, now each line gets separated into 'Charans'. Again, as same as line separation here is also some default delimiter for separation which are (",", "।", "॥") comma and pipe characters which again can be changed or modified if required.

As per the standards we choose "\n" for new lines separation and (",", "।", "॥") comma and pipe characters for the separation of 'Charans' as per the majority of 'Chhands' we deal with follows the same. After separation of the 'Charans', now each 'Charan' gets chopped into a set of words. Each word further chopped into single-single characters.

Now each character is separated and for each 'charan's' each character and diacritic the specific quantities are defined in the manner of the sequence or as per the order according to the 'Matra Gadana' or Quantities Calculation rules of 'Chhands'.

Along with that the 'Varna' Count and the sequence of the 'Varna' are also stored for the further uses in the specific 'Chhand' detection related calculation or conditions checking.

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

After the basic calculation of quantities and varnas now these values are passed into the specific core prosody conditions and calculations if it any of 'Chhand' get detected then it's metadata such as type, subtype, name, 'charans', 'charan' count, lines, line count, words count, character count etc. are displayed by automatic metadata generator in a specific managed format with proper output formatting.

The simple step by step explanation of this methodology can be understood better with the upcoming algorithm steps of this technique of metadata generation.

Method to Generate Metadata Step by Step:

Step 1. : Start
Step 2. : Input of Any Poem's 'Chhand' Data
Step 3. : Preprocessing and cleaning of Input Data
Step 4. : Separation of Lines
Step 5. : Separation of 'Charans' from each separated line.
Step 6. : Chopping 'Charans' into words.
Step 7. : Further chopping of words into Characters.
Step 8. : Matra Calculation based on the prosody rules of 'Matra Gadana' and store 'Charan' wise 'Matra' Calculations and Sum of 'Matras'.
Step 9. : 'Varna' Count based on the mechanism of counting of 'Varna' or characters. Store the sequence of 'Varna' based on the 'Matra' Count.
Step 10. : Now Check with all the specific individual rule of the specific set of 'Matrik Chhands', if found in 'Matrik Chhands' set output along with Type, Subtype and Detected 'Chhand' Name. Go to Step 13.
Step 11. : If not found in 'Matrik Chhands' than again check with all the specific individual rule of specific 'Varnik Chhands', if found in 'Varnik Chhands' set output along with Type, Subtype and Detected 'Chhand' Name. Go to Step 13.
Step 12. : If not found in 'Matrik Chhands' and 'Varnik Chhands' set of rules both than set output with the type of 'Mukt / Muktak Chhand'.
Step 13. : Print / Display the output with the help of appropriate output format along with all other relevant metadata information related to Inputted Data.
Step 14. : Step 14: STOP

With this methodology, we can cover all the major aspects of the automatic metadata generation for Hindi poetry based on the 'Chhands'. This methodology covers Detection of 'Chhand' Type, Sub Type, Classification, Sub Classification, Line and Character counts along with 'Matra' Count and 'Varna' Count based on 'Ganas'.

Along with all the above thing, there are special cases integration for 'Muktak' such as if the user enters a story instead of a poem or a part of a poem than on the basis of the sentences, the metadata generator generates metadata along with the information that the entered input seems to be a story. Apart from that, there are several other cases such as 'Aarties', 'Chalisas', and similar types of freestyle or mixed style 'Chhand' creation which directly comes under 'Muktak' only, but this metadata generator tries to process those as well and provide an appropriate output on the basis of given input type.

VII. Results

In this research work, we tested this metadata generator on 3026 inputs which include different poems, part of poems which was covering more than 30 particular 'Chhands' along with covering their classification and subclassification too. It covers the complete classification as shown in figure 1.

The result of this research work is sufficient enough to prove the robustness of this metadata generator's methodology and technical mechanism that we achieved 98.09% success ratio along with 1.91% failure due to some poor formatted text and mistakes and absence of delimiter or the over and irregular uses of delimiters.

VIII. Discussion

The results of any research work totally depend on all the inputs provided to it in a systematic way, here also this technique as explained in the methodology works well if proper inputs are provided. Even if we perform some basic preprocessing tasks on the provided Input, the automatic metadata generator expects that the input is correct as per the grammatical rules and regulation such as the uses of Diacritics, Half Characters, Joint Characters, Punctuation marks, Ardh Viram or Comma, Purn Virama or Full Stops and etc.

Checking Grammatical aspects in preprocessing can make this slightly more complex as of now, in near future research specifically on grammatical correction can be done or with the collaboration with existing grammatical correction research can give this work more strength and more effective results can be seen.

IX. Conclusion

This research work is very unique, first of its kind in world and one of the most challenging in the field of Metadata generation for Hindi poetries. It will open a new sub-domain of research possibilities under the computational linguistics research domain.

X. Acknowledgements

The first author would like to thank Symbiosis Institute of Computer Studies and Research (SICSR), Pune, India where his supervisor and second author of this paper is working.

References

[1] "Hindi - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Hindi. [Accessed: 07-Mar-2019]

[2] Contributors to Wikimedia projects, "Hanuman Chalisa - Wikipedia," Wikimedia Foundation, Inc., 27-Nov-2005. [Online]. Available: https://en.wikipedia.org/wiki/Hanuman_Chalisa. [Accessed: 07-Mar-2019]

[3] A. Kulkarni and G. Huet, Sanskrit Computational Linguistics: Third International Symposium, Hyderabad, India, January 15-17, 2009. Proceedings. Springer Science & Business Media, 2008 [Online]. Available: https://books.google.com/books/about/Sanskrit_Computational_Linguistics.html?hl=&id=Fk9_xb5K3DEC

[4] J. Kaur and J. R. Saini, "PuPoCl: Development of Punjabi Poetry Classifier Using Linguistic Features and Weighting," INFOCOMP, vol. 16, no. 1–2, pp. 1–7, Dec. 2017 [Online]. Available: http://infocomp.dcc.ufla.br/index.php/INFOCOMP/article/view/546. [Accessed: 12-Apr-2019]

[5] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles, and X. Zhang, "Rule-based word clustering for document metadata extraction," in Proceedings of the 2005 ACM symposium on Applied computing - SAC '05, Santa Fe, New Mexico, 2005, p. 1049 [Online]. Available: http://portal.acm.org/citation.cfm?doid=1066677.1066917

[6] X. Yu et al., "Using Automatic Metadata Extraction to Build a Structured Syllabus Repository," in Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 2007, pp. 337–346 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-77094-7_43. [Accessed: 12-Apr-2019]

[7] M.-T. Sagri and D. Tiscornia, "Metadata for content description in legal information," in 14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings., Prague, Czech Republic, pp. 745–749 [Online]. Available: http://ieeexplore.ieee.org/document/1232110/

[8] J. L. Klavans et al., "Computational linguistics for metadata building (CLiMB): using text mining for the automatic identification, categorization, and disambiguation of subject terms for image metadata," Multimed. Tools Appl., vol. 42, no. 1, pp. 115–138, Mar. 2009 [Online]. Available: https://link.springer.com/article/10.1007/s11042-008-0253-9. [Accessed: 12-Apr-2019]

[9] J. Kaur and J. R. Saini, "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms," in Proceedings of the 9th International Conference on Machine Learning and Computing, 2017, pp. 1–5 [Online]. Available: http://dl.acm.org/citation.cfm?id=3055635.3056589. [Accessed: 12-Apr-2019]

[10] "[No title]." [Online]. Available: https://ia601900.us.archive.org/24/items/in.ernet.dli.2015.478241/2015.478241.Chhand-sarawali.pdf. [Accessed: 09-Mar-2019]

[11] "*छंद:प्रभाकर*," Jagannathprasad Bhanu, 1935. [Online]. Available: https://ia801605.us.archive.org/15/items/in.ernet.dli.2015.322488/2015.322488.99999990253040.pdf. [Accessed: 09-Mar-2019]

[12] "*छंद - विकिपीडिया*." [Online]. Available: https://hi.wikipedia.org/wiki/%E0%A4%9B%E0%A4%82%E0%A4%A6. [Accessed: 09-Mar-2019]

[13] "*भारतीय छंद विधान*." [Online]. Available: http://www.openbooksonline.com/groups/group/show?groupUrl=chhand&%2F=. [Accessed: 09-Mar-2019]

[14] "*छन्द - भारतकोश, ज्ञान का हिन्दी महासागर*." [Online]. Available: http://bharatdiscovery.org/india/%E0%A4%9B%E0%A4%A8%E0%A5%8D%E0%A4%A6. [Accessed: 09-Mar-2019]

[15] "Kavyakala." [Online]. Available: https://kavyakala.blogspot.com/. [Accessed: 09-Mar-2019]

[16] R. Shukla, "*स्वस्थ हिन्दी समाज*." [Online]. Available: http://hhindisamaj.blogspot.com/. [Accessed: 09-Mar-2019]

[17] Hitishere et al., "*हिन्दी साहित्य काव्य संकलन – हिन्दी की कविता, लघुगीत, बालगीत, फिल्म गीत, गजल, शायरी, धार्मिक लोकगीत का विशाल संकलन*." [Online]. Available: http://www.hindisahitya.org/. [Accessed: 09-Mar-2019]

[18] "Hindi Chhandolakshan," Google Books. [Online]. Available: https://books.google.com/books/about/Hindi_Chhandolakshan.html?id=nIht8V7Xlu8C. [Accessed: 09-Mar-2019]

[19] Contributors to Wikimedia projects, "*वैदिक छंद - विकिपीडिया*," Wikimedia Foundation, Inc., 24-Dec-2014. [Online]. Available: https://hi.wikipedia.org/wiki/%E0%A4%B5%E0%A5%88%E0%A4%A6%E0%A4%BF%E0%A4%95_%E0%A4%9B%E0%A4%82%E0%A4%A6. [Accessed: 09-Mar-2019]

[20] "The Unicode Standard, Version 12.0 - Devanagari," 2019. [Online]. Available: https://www.unicode.org/charts/PDF/U0900.pdf. [Accessed: 09-Mar-2019]