Word by Word Labelling of Romanized Sindhi Text by using Online Python Tool

Irum Naz Sodhar¹
Post-Doctoral Fellow, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University Malaysia

Abdul Hafeez Buller²
Post-Doctoral Fellow, Department of Civil Engineering, Kulliyyah (Faculty) of Engineering, International Islamic University Malaysia

Suriani Sulaiman³
Assistant Professor, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University Malaysia

Anam Naz Sodhar⁴
Postgraduate Student, Quaid-e-awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan

Abstract—Sindhi is one of the most ancient languages in the world and it has its own written and spoken scripts. After the rigorous study it was found that a lot of research work has been done in different languages, but word by word labelling of Sindhi language had not been done yet. In this research study, word labelling was done on 100 sentences of Romanized Sindhi texts using Python online tool. The dataset was collected from different sources which include Sindhi newspaper, blogs and social media webpages. From this dataset, a rule-based model has been applied for the Parts-of-Speech (POS) tagging of the Romanized Sindhi sentences. A total of 624 words of Romanized Sindhi texts were tested and successfully tagged by the SindhiNLP tool in which 482 words were tagged as nouns and pronouns, 92 words tagged as verbs and 50 words tagged as determinants.

Keywords—Romanized sindhi; word labelling; rule-based model; POS tagging; SindhiNLP tool

I. INTRODUCTION

Sindhi is one of the most ancient languages in the world which has its own script in written and spoken forms [1-3]. Communication technologies are increasing day-by-day for different purposes, while different applications and software are used for daily communications such as WhatsApp, Facebook, Twitter, Telegram and Instagram [4-5]. In the community that uses Sindhi as their main language, Romanized Sindhi texts are used in daily communication especially in writing text messages on mobile phones, WhatsApp and other social media platforms [6].

Natural Language Processing has a vital role in the field of machine learning. This field provides language processing tasks such as of Parts–of–Speech tagging, tokenization of text (i.e., words, sentences, and paragraph) to the users [7-8]. In this research study, 100 sentences of Romanized Sindhi texts were labelled. The word labelling process which consists of two natural language processing tasks which is tokenization and POS tagging was performed using an online SindhiNLP tool [9]. Before performing the two tasks, a rule-based model has been applied for the POS tagging of the sentences to improve the accuracy of the POS tags [10-12].

After the review of the literature it was observed that a lot of vacuum is still available for the Sindhi language. This research study presents the word by word labelling of Sindhi language after Romanization.

II. METHODOLOGY FOR LABELLING OF ROMANIZED SINDHI

The procedure for labelling of the Sindhi Romanized text has been divided into various stages as shown in Fig. 1. The first phase involves the data collection process from different sources of Sindhi scripts, the second stage is the conversion of Sindhi scripts into Romanian scripts (i.e., Romanization), the third stage identifies the issues in word labelling after applying the rule-based model and the final task is to do a thorough analysis on the results produced [13].

A. Dataset of Sindhi Text

Sindhi language is one of the oldest, historical and most commonly used languages in the world. Sindhi language is more difficult than other languages due to the difficulty in reading, writing and understanding the scripts [13-14]. Sindhi language is spoken by the people in the province of Sindh which is the second largest populated province of Pakistan. Sindhi is the official language of the Sindh province in which almost 15% of the population use Sindhi as their mother tongue [14-15]. As Sindhi language is mostly used in Sindh-Pakistan, the data for this research study was collected within the province of Sindh. Data was collected from different sources (Sindhi newspaper, blogs, and social media webpages) which provided the rules and guidelines of Romanized Sindhi for text communication.

B. Sindhi Alphabet

Sindhi language has its own script and written style like other languages (Arabic, Urdu, and English) [16]. In Sindhi script there are 52 alphabetical letters for writing and speaking purposes and presented in Fig. 2. Sindhi language has one of the largest numbers of alphabetical letters as compared to other languages. Similar to Arabic and Urdu scripts, the Sindhi script is written from right to left with a total of 52 alphabets [17].
C. Romanization of Sindhi Text

In this research study, 100 sentences were used for the word labelling of Sindhi texts. After the collection of Sindhi sentences for the data set for this research study, the collected dataset was converted from Sindhi scripts into Romanized Sindhi text by using rules for Romanization of Sindhi text. Romanization of Sindhi text was successfully done following the rules for Romanized Sindhi text.

III. PRE-PROCESSING OF ROMANIZED SINDHI

Pre-processing is the basic components of NLP to filter the raw the data to useful and remove unnecessary data from the text. The pre-processing step consists two steps first is performing tokenization and second one assigning tag on each token [10].

A. Tokenization of Romanized Sindhi Text

The tokenization of Romanized Sindhi text has been done using the online SindhiNLP Python tool [9]. The Romanized texts were prepared following the rules of Sindhi on 100 sentences. The statistical information after the tokenization process of the Sindhi text is shown in Table I. This table consists of five different columns which are: total number of sentences, total number of words, total number of characters (with space), total number of character (without space) and total number of word tokens. In this table, two types of sentences were used: sentences from Sindhi text and sentences from Romanized Sindhi text. A total of 652 words, 2,816 characters with space and 2,262 characters without space were extracted as shown in below Table I.

| Description               | Total number of sentences | Total number of words | Total number of characters (with space) | Total number of characters (without space) | Total number of word tokens |
|---------------------------|---------------------------|-----------------------|----------------------------------------|------------------------------------------|-----------------------------|
| Sentences in Sindhi scripts | 100                       | 652                   | 2816                                   | 2262                                     | ---                         |
| Sentences in Romanized Sindhi | 100                       | 624                   | 3275                                   | 2740                                     | 624                         |

B. Parts-of-Speech Tagging

The POS tagging task for Sindhi was designed such that the whole process was divided into a few steps. The first step involved the pre-processing of the Romanized Sindhi sentences. Subsequently, the ruled-based model of Sindhi was applied for the Romanization process as described in Table II. This Romanized Sindhi text was then used as input to the SindhiNLP tool, after the input Romanized Sindhi text, the text was pre-processed using the online SindhiNLP Python tool [9] in which the sentences were split into words (i.e., tokenization). Next, the Match step was performed which was also subdivided into two categories: Assigned Tag and Incorrect Tag. If the tag was incorrectly assigned, we apply the rule-based model and repeat the process again.

C. Algorithm for POS Tagging of Romanized Sindhi Text

The algorithm for the Parts of speech tagging of Romanized Sindhi text was designed before the start of the research work. The algorithm used was based on the ten steps described below. The same step applies following the algorithm for every new input data of Romanized Sindhi text.

Step 0        Start
Step 1        Take input sentence
Step 2  Split text → words
Step 3  Repeat steps 2→7 when ≥ get appropriate output
Step 4  If word is matched, continue to assign tag separately, word by word
Step 5  If same tag is assigned to multiple words, apply rules for words and assign one tag for each word
Step 6  If one tag is assigned to one word, display the word with tag
Step 7  Else, select one or more morphological rules and apply to words to extract word with appropriate tag.
Step 8  Display as output the tagged words
Step 9  Apply rules for new words when entered
Step 10 End

D. Rule-Based Model for Labelling of Romanized Sindhi Text

The rule-based model used in the word labelling of Sindhi text is a supervised machine learning model or hybrid model. This model combines the use of online and manual approach. This type of model is commonly used to create rules for language analysis and is a popular NLP technique to perform different tasks on different languages as it is easier to understand while the results are based on ground truth values [19-20]. Fig. 3 illustrates that the S1, S2, S3 until Sn are input sentences while R1, R2, R3 until R10 are the rules. These rules are applied on the Sindhi sentences to get the appropriate output, Y. The rules for Romanized Sindhi texts are described in Table II.

![Rule Based Model](image)

Fig. 3. Rule-Based Inputs and Output for Sindhi POS Tagging

There are ten rules that have been created for the word labelling of Romanized Sindhi texts [21-22]. Rule 1 describes the structure of a sentence and the restructuring of an input sentence by applying the SVO structure (Subject + Verb + Object) [18]. Rule 2 is used to define the prefixes of Sindhi sentences (i.e., ma, mounkhe, huwa, manhon, na, wanu sijh, cha, eho, kethe, Ali, Sara etc.) as starting words and refers to nouns. Rule 3 describes the prefix that appears in sentences (i.e., he) as an initial word which is considered as pronouns. Rule 4 is used for the words that appear at the beginning of input sentences (i.e., Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara, he, etc.), considered as nouns as well as pronouns. Rule 5 describes the words that appear in the middle of an input sentence (i.e., Sadyo, Parhyo, maryo, likhyo, budho, khedan) known as the verb class. Rule 6 is used when the infix letters (i.e., a, d, e, and o) appear in between words in a sentence which refers to a verb class. Rule 7 is used for postfix letters (i.e., e, o, n, i, u), if they appear in the middle of a word in a sentence which refers to a verb class. Rule 8 is used for the postfix letters (i.e., d, e, h, o, and y) if they appear at the end of the final word in a sentence, which belongs to a noun class. Rule 9 applies when the part-of-speech tagger fails to identify when the input sentences are interrogative. Rule 10 is used when the parts-of-speech tagging is performed on sentences with negation (without subject in the sentence), otherwise it was not identified. The rules used for Romanized Sindhi Text help in performing POS tagging on the SindhiNLP tool [9] to produce a more accurate part-of-speech.

| Rule # | Rule Description | Related Examples |
|-------|------------------|------------------|
| 1     | Sentence structure should be built by applying the SVO (Subject + Verb + Object) structure. | You are teacher |
| 2     | Prefixes (Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara etc.) in sentences as starting words, refers to noun class. | I am a Student Ma/NNP shagrid/JJ |
| 3     | Prefix (he) in sentences as starting words, refers to pronoun class. | He is intelligent He/PRP ahy/VBD |
| 4     | Prefixes (Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara, he etc.) in sentences as starting words, refers to noun as well as pronoun class. | I play game Ma/NNP khedan/VBD |
| 5     | Infixes (Sadyo, Parhyo, maryo, likhyo, budho, khedan etc.) | I wrote article Ma/NNP likhyo/VBD |

TABLE II. RULES FOR ROMANIZED SINDHI TEXT FOR POS USING THE TEMPLATE
that appear in the middle of sentences, known as verb class.

| #   | Sindhi Sentence | English Sentence | Romanized Sindhi | Word Tokens | Word Labelling |
|-----|-----------------|------------------|------------------|-------------|----------------|
| 01  | اهي ڀلي هتي هضم            | They better work now | Ehe kamu kan bhole hanne | Ehe kam kan bhole hanne | Ehe/NNP kamu/VBD kan/NN bhole/NN hane/NN |
| 02  | ايني ڀلي ڀلي آرام          | We should rest now | Aseen kryon bhole hanne aram | Aseen kyon bhole hanne aram | Aseen/NNP kyon/VBD bhole/NN hane/NN aram/NN |
| 03  | ڀلي هڪ پاسمند          | She was a doctor | Huoa huehek doctor | Huoa huehek doctor | Huoa/NNP hue/VBD doctor/NN |
| 04  | اهي ڀلي ڪراچيء مهين      | I was in karachi | Maan huos Karachi maen | Maan huos Karachi maen | Maan/NNP huos/VBD Karachi/NP maen/NN |
| 05  | ڀلي ڪراچيء هڪ ڇىڪري       | You are a handsome boy | Tawan huoao hek khubhsorat chokra | Tawan huoas hek khubhsorat chokra | Tawan/NMP hua/VBD hek/NN khubhsorat/NN chokra/NN |
| 06  | ڀلي ڪراچيء هڪ ڇىڪري       | She was an attractive girl | Huoa hui hek mohendar chokri | Huoa hui hek mohendar chokri | Huoa/NNP hua/VBD hek/NN mohendar/NN chokri/NN |
| 07  | اهي ڌڪوندڙڙر هر             | It was painful | Eho dukhioindar ho | Eho dukhioindar ho | Eho/NNP dukhioindar/NN ho/NP |
| 08  | ايني ڀلي افس مهين              | We were here in the office | Aseen huos seen office maen | Aseen huos seen office maen | Aseen/NNP huos seen/VBD he/NN office/NN maen/NN |
| 09  | ڀلي پهرين سال واري           | He was in the first year class | He ho pehreyen saal ware class maen | He ho pehreyen saal ware class maen | He ho pehreyen saal/VBD saal/NN ware/NN class/NN maen/NN |
| 10  | اهي ڪمعن راد ڪي مينا          | They were in the playground yesterday | Uhe huoa rand maidan mean kalh | Uhe huoa rand maidan mean kalh | Uhe/NNP huoa/NN rand/NN maidan/NN mean/NN kalh/NN |

IV. WORD BY WORD LABELLING OF ROMANIZED SINDHI

Word labelling of Romanized Sindhi Text was performed using the free online SindhiNLP Python tool [9]. Word Labelling of Romanized Sindhi text has been performed after completing the two pre-processing tasks for Sindhi Romanized text: the tokenization and part-of-speech tagging tasks as shown in Table III.
A. **Analysis of the Parts-of-Speech Tagging**

The output from the word labelling task of Romanized Sindhi text performed using the online SindhiNLP Python tool [9] and Sindhi rule-based model was analyzed in which 13 different POS categories were identified. The detailed statistics of the word labelling task are shown in Table IV.

| Description                  | Total Number of Words | Total number of POS Tagged Words | Word Labelling of Romanized Sindhi Text |
|------------------------------|-----------------------|----------------------------------|----------------------------------------|
| Romanized Sindhi Text (100 sentences were used) | 624                   | 624                              |                                        |
| POS                         | No. of Words          |                                  |                                        |
| NNP                         | 110                   |                                  |                                        |
| NN                          | 372                   |                                  |                                        |
| PRP                         | 11                    |                                  |                                        |
| JJ                          | 13                    |                                  |                                        |
| RB                          | 11                    |                                  |                                        |
| WP                          | 4                     |                                  |                                        |
| VBD                         | 54                    |                                  |                                        |
| VBZ                         | 0                     |                                  |                                        |
| VBN                         | 4                     |                                  |                                        |
| VBP                         | 30                    |                                  |                                        |
| VB                          | 4                     |                                  |                                        |
| WDT                         | 0                     |                                  |                                        |
| DT                          | 11                    |                                  |                                        |
| Total                       | 624                   | 624                              |                                        |

From the results produced by the SindhiNLP POS tagger, 624 Sindhi words was successfully tagged in which 482 are noun and pronouns, 92 verbs and 50 determinants were found as illustrated in Fig. 4.

V. **Conclusion**

In research study of Word by Word Labelling of Romanized Sindhi Text the conclusion is based on the following outcomes.

- A hybrid approach was used that combines online and manual approaches and a rule-based algorithm was designed and applied to the word labelling tasks.
- From the results 13 different POS categories were identified and 654 words of Romanized Sindhi Text were tested by using the SindhiNLP Python tool and all words were tagged successfully.
- From the results 482 noun/pronouns were found while the remaining 172 words were found to be adjectives, adverbs, verbs and determiners.
- For future work, Romanized Sindhi text from different domains will be used in the word labelling tasks and results will be compared using different machine learning techniques and tools.

**References**

[1] Iyengar, Arvind. "A diachronic analysis of Sindhi multiscriptality." Journal of Historical Sociolinguistics 7, no. 2 (2021): 207-241.

[2] J. Lalwani, "History of Sindhi Language", Voice of Sindhiastan, Vol. 4, no. 4, (2005). http://www.sindhishaan.com/article/language/lang_04_04.html

[3] History of Sindh - Govt. of Sindh [Retrieved on June 27, 2022]. https://www.sindh.gov.pk/history

[4] Nair, Jayashree, Riyaz Ahamed, and Anakha Shaji. "A Study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages." In Sustainable Communication Networks and Application, pp. 103-117. Springer, Singapore, 2022.

[5] Ali, Wazir, Rajesh Kumar, Yong Dai, Jay Kumar, and Saifullah Tumrani. "Neural Joint Model for Part-of-Speech Tagging and Entity Extraction." In 2021 13th International Conference on Machine Learning and Computing, pp. 239-245. 2021.

[6] Saeed, Hafiz Hassaan, Muhammad Haseeb Ashraf, Faisal Kamiran, Asim Karim, and Toon Calders. "Roman Urdu toxic comment classification." Language Resources and Evaluation 55, no. 4 (2021): 971-996.

[7] AL. MANSOORI, M. O. U. Z. A. "Exploring Sentiment Analysis using Different Machine Learning Algorithms on Dialectal Arabic." PhD diss., The British University in Dubai (BUiD), 2021.

[8] Arora, Gaurav. "InLTk: Natural language toolkit for indic languages." arXiv preprint arXiv:2009.12534 (2020).

[9] Online Python tool http://text-processing.com/demo/

[10] Li, Hongwei, Hongyan Mao, and Jingzi Wang. "Part-of-Speech Tagging with Rule-Based Data Preprocessing and Transformer." Electronics 11, no. 1 (2021): 56.

[11] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Parts of speech tagging of Romanized Sindhi text by applying rule based model." IJCNS 19, no. 11 (2019): 91.

[12] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Abdul Hafeez Buller, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Sentiment analysis of Romanized Sindhi text." Journal of Intelligent & Fuzzy Systems 38, no. 5 (2020): 5877-5883.

[13] Sodhar, Irum Naz, Akhtar Hussain Jalbani, and Muhammad Ibrahim Channa. "Identification of issues and challenges in romanized Sindhi text." International Journal of Advanced Computer Science and Applications 10, no. 9 (2019).

[14] Abbasi, Muhammad Hassan, and Sajida Zaki. "LANGUAGE SHIFT: JOURNEY OF THIRD GENERATION SINDHI AND GUJARATI
SPEAKERS IN KARACHI.” Bahria University Journal of Humanities & Social Sciences 2, no. 1 (2019): 19-19.

[15] Shackle, C. "Sindhi language." Encyclopedia Britannica, July 9, 2018. https://www.britannica.com/topic/Sindhi-language.

[16] Zeroual, Imad, Abdelhak Lakhouaja, and Rachid Belahbib. "Towards a standard Part of Speech tagset for the Arabic language." Journal of King Saud University-Computer and Information Sciences 29, no. 2 (2017): 171-178.

[17] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Romanized Sindhi rules for text communication." Mehran University Research Journal Of Engineering & Technology 40, no. 2 (2021): 298-304.

[18] Afini, Umriya, Catur Supriyanto, and Raden Arief Nugroho. "The Development of Indonesian POS Tagging System for Computer-aided Independent Language Learning." International Journal of Emerging Technologies in Learning 12, no. 11 (2017).

[19] Ekbal, Asif, S. Mondal, and Sivaji Bandyopadhyay. "POS Tagging using HMM and Rule-based Chunking." The Proceedings of SPSAL 8, no. 1 (2007): 25-28.

[20] Devi, S. Anjali, and S. Sivakumar. "A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets." International Journal of Advanced Computer Science and Applications 12.9 (2021).

[21] Btoush, Mohammad Hjouj, Abdulsalam Alarabeyyat, and Isa Olab. "Rule based approach for Arabic part of speech tagging and name entity recognition." International Journal of Advanced Computer Science and Applications 7.6 (2016).

[22] Khan, Sadiq Nawaz, et al. "Urdu word segmentation using machine learning approaches." International Journal of Advanced Computer Science and Applications 9.6 (2018).