Documents

Sodhar, I.N.[a] , Buller, A.H.[b] , Sulaiman, S.[a] , Sodhar, A.N.[c]

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

(2022) *International Journal of Advanced Computer Science and Applications*, 13 (8), pp. 262-267. Cited 2 times.

DOI: 10.14569/IJACSA.2022.0130831

[a] Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University Malaysia, Malaysia
[b] Department of Civil Engineering, Kulliyyah (Faculty) of Engineering, International Islamic University Malaysia, Malaysia
[c] Quaid-e-awam University of Engineering, Science & Technology, Sindh, Nawabshah, 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.

Author Keywords
Pos tagging; Romanized sindhi; Rule-based model; Sindhinlp tool; Word labelling

Index Keywords
Computational linguistics, Python; Labelings, Newspaper medium, On-line tools, PoS tagging, Research studies, Romanized sindhi, Rule-based models, Sindhinlp tool, Social media, Word labeling; High level languages

References

- Iyengar, Arvind
  A diachronic analysis of Sindhi multiscriptality
  (2021) *Journal of Historical Sociolinguistics*, 7 (2), pp. 207-241.
  [1]
- Lalwani, J.
  History of Sindhi Language
  (2005) *Voice of Sindhistaan*, 4 (4).
  [2]
- *History of Sindh - Govt. of Sindh*,
  [3] [Retrieved on June 27, 2022]
- Nair, Jayashree, Ahammed, Riyaz, Shaji, Anakha
  A Study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages
  (2022) *Sustainable Communication Networks and Application*, pp. 103-117.
  [4] Springer, Singapore
- Ali, Wazir, Kumar, Rajesh, Dai, Yong, Kumar, Jay, Tumrani, Saifullah
  Neural Joint Model for Part-of-Speech Tagging and Entity Extraction
  (2021) *2021 13th International Conference on Machine Learning and Computing*, pp. 239-245.
  [5]
- Saeed, Hafiz Hassaan, Ashraf, Muhammad Haseeb, Kamiran, Faisal, Karim, Asim, Calders, Toon

Roman Urdu toxic comment classification
(2021) *Language Resources and Evaluation*, 55 (4), pp. 971-996.
[6]

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

- Arora, Gaurav
(2020) *iNLTK: Natural language toolkit for indic languages*,
[8] arXiv preprint arXiv:2009.12534

- *Online Python tool*,
[9]

- Li, Hongwei, Mao, Hongyan, Wang, Jingzi
Part-of-Speech Tagging with Rule-Based Data Preprocessing and Transformer
(2021) *Electronics*, 11 (1), p. 56.
[10]

- Sodhar, Irum Naz, Jalbani, Akhtar Hussain, Channa, Muhammad Ibrahim, Hakro, Dil Nawaz
Parts of speech tagging of Romanized Sindhi text by applying rule based model
(2019) *IJCSNS*, 19 (11), p. 91.
[11]

- Sodhar, Irum Naz, Jalbani, Akhtar Hussain, Buller, Abdul Hafeez, Channa, Muhammad Ibrahim, Hakro, Dil Nawaz
Sentiment analysis of Romanized Sindhi text
(2020) *Journal of Intelligent & Fuzzy Systems*, 38 (5), pp. 5877-5883.
[12]

- Sodhar, Irum Naz, Jalbani, Akhtar Hussain, Channa, Muhammad Ibrahim
Identification of issues and challenges in romanized Sindhi text
(2019) *International Journal of Advanced Computer Science and Applications*, 10 (9).
[13]

- Abbasi, Muhammad Hassan, Zaki, Sajida
LANGUAGE SHIFT: JOURNEY OF THIRD GENERATION SINDHI AND GUJARATI SPEAKERS IN KARACHI
(2019) *Bahria University Journal of Humanities & Social Sciences*, 2 (1), pp. 19-19.
[14]

- Shackle, C.
Sindhi language
(2018) *Encyclopedia Britannica*,
[15] July 9

- Zeroual, Imad, Lakhouaja, Abdelhak, Belahbib, Rachid
Towards a standard Part of Speech tagset for the Arabic language
(2017) *Journal of King Saud University-Computer and Information Sciences*, 29 (2), pp. 171-178.
[16]

- Sodhar, Irum Naz, Jalbani, Akhtar Hussain, Channa, Muhammad Ibrahim, Hakro, Dil Nawaz
Romanized Sindhi rules for text communication
(2021) *Mehran University Research Journal Of Engineering & Technology*, 40 (2), pp. 298-304.
[17]

- Afini, Umriya, Supriyanto, Catur, Nugroho, Raden Arief
  The Development of Indonesian POS Tagging System for Computer-aided Independent Language Learning
  (2017) *International Journal of Emerging Technologies in Learning*, 12 (11).
  [18]

- Ekbal, Asif, Mondal, S., Bandyopadhyay, Sivaji
  POS Tagging using HMM and Rule-based Chunking
  (2007) *The Proceedings of SPSAL*, 8 (1), pp. 25-28.
  [19]

- Devi, S. Anjali, Sivakumar, S.
  A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets
  (2021) *International Journal of Advanced Computer Science and Applications*, 12 (9).
  [20]

- Btoush, Mohammad Hjouj, Alarabeyyat, Abdulsalam, Olab, Isa
  Rule based approach for Arabic part of speech tagging and name entity recognition
  (2016) *International Journal of Advanced Computer Science and Applications*, 7 (6).
  [21]

- Khan, Sadiq Nawaz
  Urdu word segmentation using machine learning approaches
  (2018) *International Journal of Advanced Computer Science and Applications*, 9 (6).
  [22]

Publisher: Science and Information Organization

ISSN: 2158107X
Language of Original Document: English
Abbreviated Source Title: Intl. J. Adv. Comput. Sci. Appl.
2-s2.0-85137164332
Document Type: Article
Publication Stage: Final
Source: Scopus