PAKISTAN JOURNAL OF EMERGING SCIENCE AND TECHNOLOGIES (PJEST)

December, 2020, 1(1) ISSN XXXX – XXXX (On-Line)

TOKENIZATION OF SINDHI TEXT ON INFORMATION RETRIEVAL TOOL

Irum Naz Sodhar[1,*], Akhtar Hussain Jalbani[2], Abdul Hafeez Buller[3]& Anam Naz Sodhar[4]

[1]Lecturer, Department of Information Technology, Shaheed Benazir Bhutto University, Shaheed Benazir Abad, Sindh-Pakistan. Email: irumnaz@sbbusba.edu.pk

[2] Associate Professor, Department of Information Technology, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah, Sindh-Pakistan. Email: jalbaniakhtar@gmail.com

[3] Engineer, Engineering Section, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah, Sindh-Pakistan. Email: ah.buller@quest.edu.pk

[4]Postgraduate Student, Quaid-e-Awam University of Engineering Science & Technology, Nawabshah Sindh-Pakistan. Email: anumakber10@gmai.com

**Correspondence:* irumnaz@sbbusba.edu.pk

ARTICLE INFORMATION

Citation: S. Irum Naz, J. Akhtar Hussain, B. Abdul Hafeez, and S. Anam Naz, "TOKENIZATION OF SINDHI TEXT ON INFORMATION RETRIEVAL TOOL" PJEST, vol. 1, p. 7, 7 May 2021.

http://doi.org/10.5281/zenodo.4774104

Received: 15th-April-2021

Revised and Accepted:

06th -May-2021

Published On-Line:

07th -May-2021

***Corresponding Author:**

Irum Naz Sodhar:

irumnaz@sbbusba.edu.pk

Original Research Article

ABSTRACT

In the artificial intelligence divided into sub fields, Natural language processing (NLP) is also field of AI and performs lot of NLP task on scripts. Tokenization is also important task of NLP to break the text. Tokenization process used to text identifies their text and text count. In this research study focus on tokenization to perform task on Sindhi sentences by using tool and get information retrieval from tool. Corpus used Awami newspaper of Sindhi on the basis of sentence form. Information retrieval based on tool's response and also helps users to in Simplification, satisfaction, filtration of text and so on. Tokenization task considered as pre-processing task of NLP and produce tokens with token count which is the basis on given input text to information retrieval Tokenization tool. One hundred forty words of Sindhi text and eight sentences were used to get results. In future, perform NLP tasks on Sindhi text by using supervised, Semi-supervised and unsupervised machine learning.

Keywords: Artificial Intelligence, Natural Language Processing, Sindhi, Tokenization, Information Retrieval tool.

Introduction:

Sindhi is one of the oldest languages and are spoken in all over the world, but especially in the Sindh-Pakistan [1-3]. Sindhi language is read write and speak in the Sindh province. Today's technology is increasing rapidly throughout the world almost in every field of life. Information technology is one of them which is growing rapidly prove easiness to the users. Artificial

Intelligence plays vital role in Information Technology/Computer Science for various sub-fields such as the Natural Language Processing is one of them [4].

a) Natural Language Processing:

Natural Language Processing (NLP) is the most important part of area in artificial intelligence that supports machines to comprehend and manipulate human speaking/ writing language. NLP used in lots of fields to fill the research gap in between machine and human communication such as: Email, Messages, Letter, Fax and Question/ Answering [5]. The area of NLP two important parts are included such as: one is Speak form (for speaking is an important and basic way communications to humans and understands each other) second is Written form (for written is also important way to communications and officially used) [6].

b) Sindhi Language:

Sindhi Language (SL) is an oldest Language of the world. History of Sindhi language is five thousand years ago. This language having fifty-two letters in alphabet with different placement of dots just like: Above, Below in between. Sindhi Language is grammatically complex and rich is morphologically [7-8]. Sindhi and English Languages are different from grammatically structure and sense of recognition [9]. This language is used in almost world widely as written, read and spoken by people [10].

c) Information Retrieval:

Information Retrieval (IR) is continuously interest in research and lot of chances in the area of data mining. An IR focuses with illustration, storing, access and retrieve information regarding to users input [11-12]. Following are some recent research ideas [13] in the field of IR such as: Information Searching (IS), Ranking/Indexing (R/I) of user's inputs result, elaborating representation (ER) and storage of information (SI), Classification of documents (Pre-defined groups), Clustering of documents (Automatically creates clusters). Information Retrieval (IR) main focus on to identify of tokens.

d) Information Retrieval Tool:

Information Retrieval Tool (IRT) is used as SindhiNLP which is developed by Mazhar Ali Dootio to solve the problems of Computational Sindhi Language (SL) and perform NLP task on this tool and retrieve easily information on this tool [14-18]. The tasks perform on this tool just as: Sindhi Online text Parser (SOTP), Sindhi WordNet (SWN), Sindhi Lemma (SL), Sindhi Stemmer (SS) and Sentiment Analysis (SA) by using input Sindhi Text.

Methodology:

This research study is depending on four major steps such as: (a) input Text, (b) Processing on tool (IRT), (c) Tokenization of input text (d) Retrieve information from tool. Fewer of minor

steps just as: break sentences into word form, Tokens in word form, Tokens were in numerical numbers etc.

Fig.1 shows the methodology of research to perform per processing task on Sindhi text by using tool (IRT). Above steps help to perform tokenization and get appropriate result from tool.

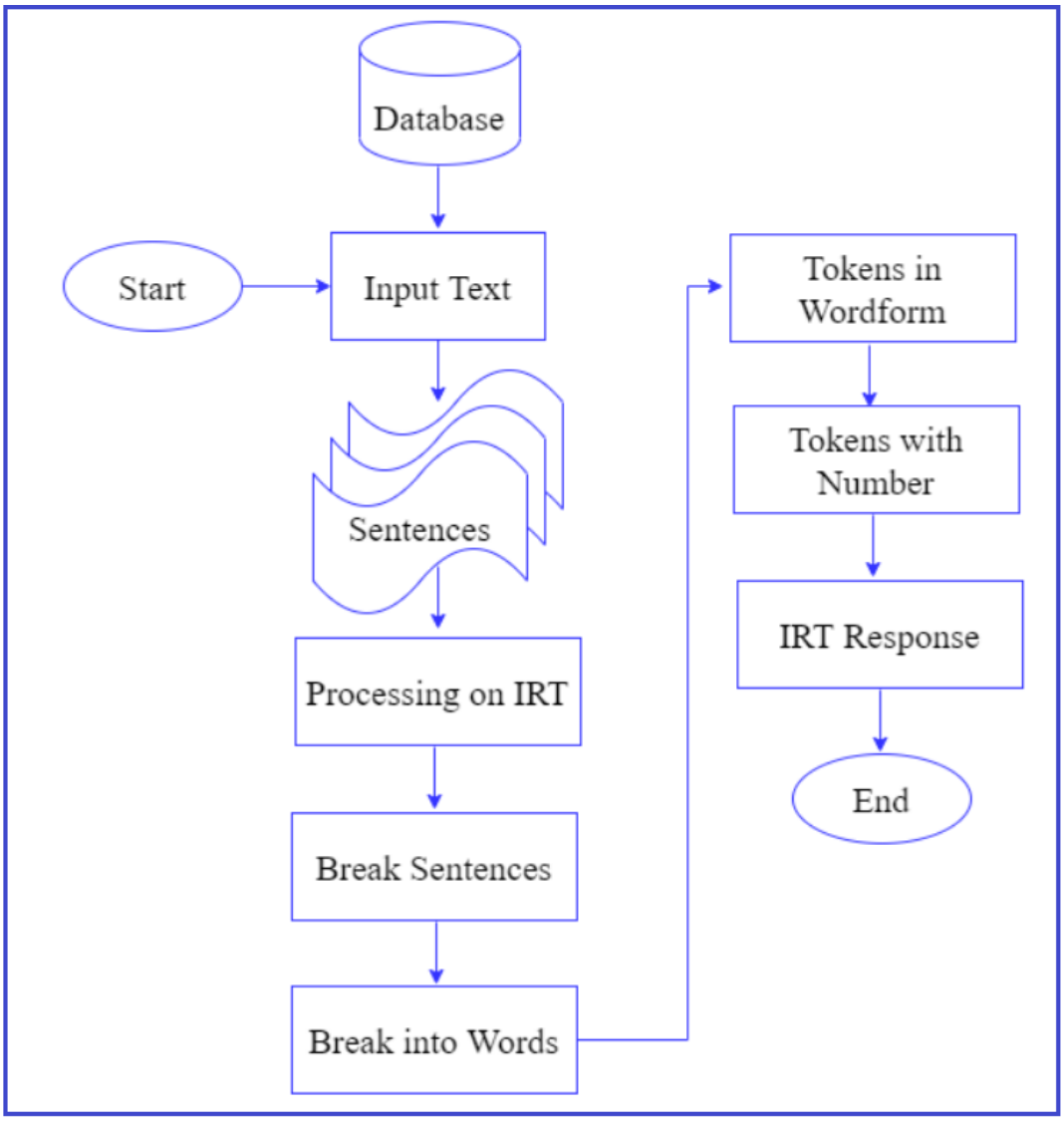


Fig. 1: Methodology of Research

a) Corpus:

In this research study corpus used as Sindhi text in the form of sentences. Eight Sentences with one hundred forty words were used. This corpus based on Sindhi Language that is mostly used in Sindh Province of Pakistan. This language contains fifty-two 52- alphabetic letters with dots and writing style used as right-handed language. Those corpuses contain Sindhi text contain one twenty-six words with fourteen-14 numerical words in corpus. That corpus takes from online

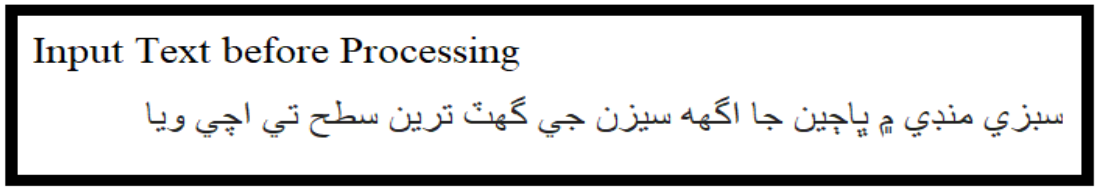
Awami Sindhi newspaper from 6th January 2021 to perform NLP pre-processing task on IRT tool. Table 1 shows the research corpus and also used in this research.

Table I: Corpus of Research

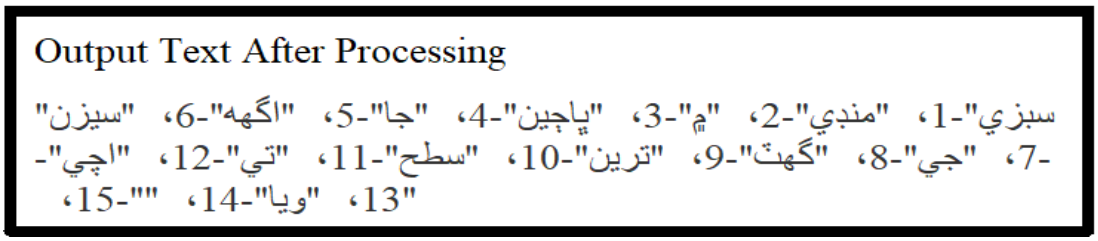
Sindhi Sentences	S. No.
سبزي منڊي ۾ پاڇين جا اڳهه سيزن جي گهٽ ترين سطح تي اچي ويا	1
سبزي منڊي مارڪيٽ ڪميٽي موجب سبزي منڊي ۾ پڻاٽا ۽ بصر 25 رپين ۾ وڪرو ٿي رهيا آهن	2
سبزي منڊي ۾ ٽماٽا25 رپيا پر شهر ۾ 100 رپيا في ڪلو وڪرو ٿي رهيا آهن	3
ڪيرا منڊي ۾ 25 شهر ۾ 60 رپيا ڪلو، گجرن جو پرپور فصل لهڻ کان پوءِ قيمت 30 رپيا ڪلو تي اچي وئي	4
جڏهن ته شهر ۾ بصر جي قيمت 40 کان 50 رپيا وصول ڪيا پيا وڃن	5
پر شهر ۾ گجر 50 رپيا في ڪلو وڪرو ٿي رهي آهي	6
سنڌ ۾ مٽر منڊي ۾ 65 پنجاب جي منڊي ۾ 55 رپيا وڪرو ڪيا ويا پر شهر ۾ مٽر 100 رپيا ڪلو وڪرو ڪيا پيا وڃن	7
ساوا مرچ منڊي ۾ 40 رپيا ۽ شهر ۾ 200 رپيا ڪلو وڪرو ٿي رهيا آهن	8

b) Tokenization:

This research study is based on tokenization [19-20]. Tokenization is the pre-processing step of NLP field. The word tokenization is the method of break down text into words form with count numbers. This is the basic step of NLP to perform and identify the text with word count as shown in Fig. 2 (a)&(b).



(a) Input Text Processing



(b) Output Text Processing

Fig. 2: Text Processing Before & After

Results and Discussion:

This research study performs pre-processing task of tokenization on Sindhi text by using Information Retrieval System. Total eight sentences were used on Information Retrieval System and performed the task of tokenization of each sentence. Sentence-1 having fifteen words and all words were tokenized by processing on Information Retrieval tool and obtained the fifteen tokens. Sentence-2 having eighteen words with one numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of eighteen tokens. Sentence-3 having sixteen words with two numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of sixteen tokens. Sentence-4 having twenty-three words with four numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of twenty-three tokens. Sentence-5 having fifteen words with two numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of fifteen tokens. Sentence-6 having twelve words with one numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of twelve tokens. Sentence-7 having twenty-six words with three numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of twenty-three tokens and seven words did not break into token and nor show words as an output. Sentence-8 having sixteen words with two numeric number and all words were tokenized by processing on Information Retrieval tool and obtained the results of sixteen tokens. This experimental study of tokenization was performed on all eight sentences correctly but sentence-7 did not showed seven tokens of sentence when give input to information retrieval tool result.

Conclusion & Future work:

In this research study based on Sindhi text in the sentence form to perform pre-processing task of Natural Language Processing (NLP) on Information Retrieval tool (SindhiNLP). The result shown in the paper are based on the experimental over more than hundred input words in the sentences form. When given input one to twenty words perform tokenization properly and increase number of words from more than twenty-five results shows missing tokens from text on Information Retrieval tool. This research study is helpful for the researcher to perform more natural language processing task on Sindhi text by using different technique. Future work, still need to improve tool by using supervised algorithms.

Abbreviations:

AI	Artificial-Intelligence
NLP	Natural-Language-Processing
IRT	Information-Retrieval-tool
SL	Sindhi Language

SOTP	Sindhi Online text Parser
SWN	Sindhi WordNet
SL	Sindhi Lemma
SS	Sindhi Stemmer
SA	Sentiment Analysis
IR	Information Retrieval
IS	Information Searching
R/I	Ranking/Indexing
ER	Elaborating Representation
SI	Storage of Information

Author's Contribution: I.N.S., Conceived the idea; A.H.J., Designed the simulated work; A.N. S., did the acquisition of data; I.N.S., & A.N.S., executed simulated work, data analysis or analysis and interpretation of data; I.N.S., A.H.B., wrote the basic draft and did the language and grammatical edits or Critical revision.

Funding: The publication of this article was funded by no one.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgement: The corresponding author and co-authors would like to thanks the Dr. Akhtar Hussain Jalbani for assistance such kind of research and designed the simulated work.

REFERENCES

[1] W. Ali, N. Ali, and S. Tumrani, "Creating and Evaluating Resources for Sentiment Analysis in the Low-resource Language : Sindhi," pp. 188–194, 2021.

[2] Q. Talpur, I. Kakepoto, and K. B. Jalbani, "Engineering Students Perceptions about English Language Teachers Code Switching from English to Sindhi Language Engineering Students Perceptions about English Language Teachers Code Switching from English to Sindhi Language," no. April, 2021.

[3] Z. Bhatti, I. A. Ismaili, W. J. Soomro, and D. N. Hakro, "Word Segmentation Model for Sindhi Text," vol. 2, no. 1, pp. 1–7, 2014, doi: 10.12691/ajcrr-2-1-1.

[4] S. K. Srivastava, "Applications of Intelligent Agents," Electron. Inf. Plan., vol. 26, no. 5, pp. 273–281, 1999, doi: 10.1007/978-3-662-03678-5_1.

[5] T. R. Soomro& S. M. Ghulam, "Current Status of Urdu on Twitter," Sukkur IBA J. Comput. Math. Sci., 2019, doi: 10.30537/sjcms.v3i1.397.

[6] I. N. Sodhar, A. H. Jalbani, A. H. Buller, M. I. Channa, and D. N. Hakro, “Sentiment analysis of Romanized Sindhi text,” J. Intell. Fuzzy Syst., vol. 38, no. 5, pp. 5877–5883, 2020, doi: 10.3233/JIFS-179675.

[7] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, “Identification of issues and challenges in romanized Sindhi text,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 9, pp. 229–233, 2019, doi: 10.14569/ijacsa.2019.0100929.

[8] I. N. Sodhar, A. H. Jalbani, and A. H. Buller, “An Empirical And Statistical Study On Pos Tagging Of Sindhi Social Media Text,” vol. 241, pp. 72–81, 2020.

[9] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, “Parts of Speech Tagging of Romanized Sindhi Text by applying Rule Based Model,” vol. 19, no. 11, pp. 91–96, 2019.

[10] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, “Romanized Sindhi Rules for Text Communication,” vol. 40, no. 2, pp. 298–304, 2021, doi: 10.22581/muet1982.2102.04.

[11] G. Salton and J. McGill, Michael, “Information Retrieval: an Introduction,” in Introduction to modern information retrieval, 1983.

[12] K. Jarvelin and J. Kekalainen, “IR evaluation methods for retrieving highly relevant documents,” SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval), 2000, doi: 10.1145/3130348.3130374.

[13] B. Saini, V. Singh, and S. Kumar, “Information retrieval models and searching methodologies: Survey,” Inf. Retr. Boston., 2014.

[14] “Multimedia Based e-Learning for Educating Children in Sindhi Language,” Sukkur IBA J. Comput. Math. Sci., 2020, doi: 10.30537/sjcms.v4i1.518.

[15] I. N. Sodhar, H. Bhanbhro, Z. H. Amur, A. H. Jalbani, and A. H. Buller, “Sindhi Language Processing on Online SindhiNLP Tool,” vol. 4, no. 3, pp. 4–7, 2020.

[16] I. N. Sodhar, A. H. Jalbani, A. H. Buller, and A. N. Sodhar, “Tools Used In Online Teaching and Learning through Lock - Down,” no. 8, pp. 36–40, 2020.

[17] I. N. Sodhar, A. H. Buller, and A. N. Sodhar, “Identification of Online Statistical Translation and Text Issues in Communication Technologies,” vol. 10, no. 2, pp. 446–452, 2021.

[18] I. H. Sodhar et al., “Information Communication and Technology Tools Integration in Higher Education,” Int. J. Progress. Sci. Technol. (IJPSAT, vol. 15, no. 1, pp. 127–133, 2019, [Online]. Available: https://www.researchgate.net/publication/333984007.

[19] N. Otani, S. Ozaki, X. Zhao, Y. Li, M. St Johns, and L. Levin, “Pre-tokenization of Multi-word Expressions in Cross-lingual Word Embeddings,” 2020, doi: 10.18653/v1/2020.emnlp-main.360.

[20] C. Ding et al., “Towards Burmese (Myanmar) morphological analysis: Syllable-based Tokenization and Part-of-speech Tagging,” ACM Trans. Asian Low-Resource Lang. Inf. Process., 2019, doi: 10.1145/3325885.