University of Sindh Journal of Information and Communication Technology
(USJICT)
Volume 4, Issue 3, October 2020

ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/
© Published by University of Sindh, Jamshoro

Sindhi Language Processing on Online SindhiNLP Tool

Irum Naz Sodhar[1]*, Hina Bhanbhro[1], Zira Hassan Amur[1], Akhtar Hussain Jalbani[2]
& Abdul Hafeez Buller[3]

1 Information Technology Department, Shaheed Benazir Bhutto University, Shaheed Benazirabad Sindh, Pakistan
2 Information Technology Department, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah.
3 Engineer, Engineering Section, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah.

E-mail address: irumnaz@sbbusba.edu.pk, hina@sbbusba.edu.pk, zirahassan@sbbusba.edu.pk, jalbaniakhtar@quest.edu.pk
& engrabdulhafeezbuller@gmail.com

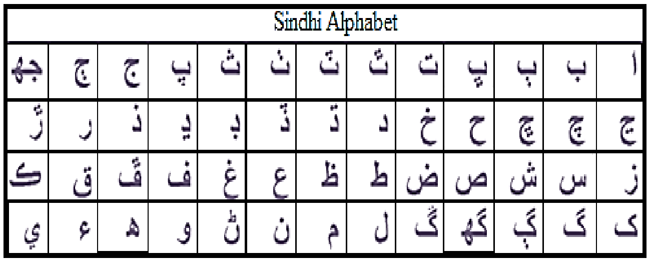
Abstract: In artificial intelligence (AI) most important field is machine learning (ML) and machine learning sub field is natural language processing (NLP) to perform different assignment by using language on machine. In this research used Sindhi language (SL) to performed tasks on SindhiNLP tool. SindhiNLP tool is freely available online and perform different tasks by using Sindhi language. This research based on seven tasks such as: (Lemma (L), Stem Suffix (SS), Stem Affix (SA), Stem (S), Sindhi Parts-of-Speech (SPOS), Universal Parts-of-Speech (UPOS) and words (W) from input sentences of Sindhi language. Number of sentences is used for research ten (10) and one ninety-five words for research to find out root words from Sindhi data. Sindhi is oldest language of world and also needs more attention for research to perform machine learning tasks by applying machine learning algorithms and tools to get results. Still no any trained data set are available for Sindhi language online.

Keywords: Artificial Intelligence (AI), Machine Learning (ML) Natural Language Processing (NLP) Sindhi Language (SL), SindhiNLP Tool

I. Introduction

Sindhi is an historical written, speaking and reading language of the world, which is mostly used by the Sindhi peoples lived in Sindh province of Pakistan. Around 12% of population of Pakistan having mother tongue is Sindhi and an officially used in Sindh [1]. Sindhi language is used in different area of world for communication either verbal or written. Sindhi language have (52 –Fifty-two) characters used in written [2], [3], [4] & [5] as shown in figure 1. Sindhi is right-handed language just like Arabic and Urdu language and follow the rules same as Arabic and Urdu language [6], [7] & [8]. Day by day increase number of users on social media so Sindhi is considered as usage of greater than before on social media such as: Online newspaper, Books, Learning Websites, Online Sindhi literature and used for communication on social media for sharing information with each other [9] & [10].

In Sindhi data build, comes problems such as data (gaining, pre-processing and tokenization) is discussed in this paper. The outcomes of those problems depend on observation which contains uni-gram, bi-gram and tri-gram frequencies, author investigate the orthography and Sindhi language build up data [11]. The word corpus used first time by German scientist. The corpora are singular of corpus and used as data set consists of huge amount of text data. Pre-processing is not easy because shortage of resource for computational linguistic and exploration, different text data have been built in different languages of different countries [1], [12]. A transliteration model provided two languages one was Perso-Arabic language and second Devanagari language. Analysis of both languages, authors recommended that data from the Roman language is also used for Sindhi language and designed algorithm for transliteration between two languages [3].



Sindhi Alphabet												
جھ	ڄ	ج	پ	ث	ٺ	ٽ	ٿ	ت	ڀ	ٻ	ب	ا
ڙ	ر	ذ	ڍ	ڊ	ڏ	ڌ	د	خ	ح	ڇ	چ	ڃ
ڪ	ق	ڦ	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز
ي	ء	ھ	و	ڻ	ن	م	ل	ڱ	گھ	ڳ	گ	ک

Figure 1. Sindhi Alphabet [9]

In another research addressed the issues of word segmentation in Sindhi and provided different method and algorithms [13]. Now-a-days research on language has lot of challenges and in Urdu and tokenize the text in different forms [14]. Arabic language also works done on sentiment analysis from un-structured and non-grammatical language [15], [16].

II. Research Methodology

This study is divided into three major steps such as: Step one contains data set and pre-process the data used Sindhi data set, Step second perform process on tool those data should be processed on tool and finally third step evaluate the results get from tool by applied as input as shown in figure 2.

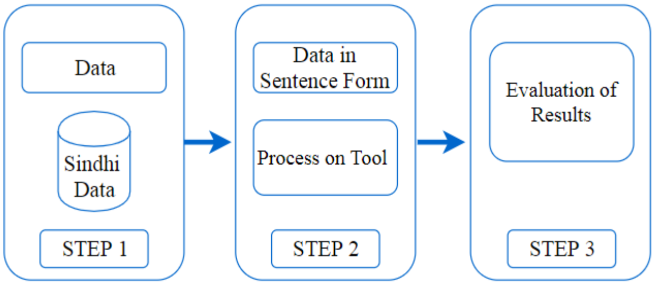


Figure 2. Research Methodology

A. Data set

Data set was collected form the Awami Newspaper as shown in table 1, which is mostly used in all over the Sindh province.

TABLE I. SINDHI DATA SET FOR EVALUATION ON TOOL

Sindhi Sentences	S. No.
سنڌ ڪابينا صوبي ۾ شاگرد يونين کي بحال ڪرڻ جي لاءِ بل منظور ڪري ورتو	1.
دنيا جي وڏي ويب سائيٽ فيس بڪ جا انجينيئر هن وقت ميسينجر ۾ هڪ نئون فيچر متعارف ڪرائڻ لاءِ سرگرم آهن	2.
ٽوئيٽر طرفان نئين پاليسي جو اعلان ڪيو ويو آهي	3.
انساني حقن جي عالمي تنظيم ايمنسٽي انٽرنيشنل گوگل ۽ فيس بُڪ جهڙين وڏين انٽرنيٽ ڪمپنين جي ڪاروباري ماڊل کي انساني حقن لاءِ خطرو قرار ڏئي ڇڏيو	4.
نوابشاهه جي معذور نوجوان سرڪاري نوڪري ڏيڻ جي اپيل ڪئي آهي	5.
انٽرنيٽ ميسيجنگ جي مشهور ترين ايپ واٽس ايپ هلندڙ سال واهپيدارن لاءِ جديد فيچرز جي فراهمي کي يقيني بڻائي ڪيترائي نوان فيچرز ايپ ۾ شامل ڪيا آهن	6.
سموري دنيا ۾ انٽرنيٽ جي فراهمي لاءِ اسپيس ايڪس ڪمپني اسٽار لنڪ جي سلسلي جي 60 سيٽلائيٽ جي کيپ ڪاميابي سان خلا ۾ روانو ڪري ڇڏي	7.
ڪمپيوٽر ۽ انٽرنيٽ ٽيڪنالوجي جي هن دور ۾ جتي تعليم ۽ صحت جي شعبي ۾ “آرٽيفيشل انٽيليجنس” مصنوعي ذهانت جو استعمال ڪيو پيو وڃي	8.
هاڻ انهيءَ ٽيڪنالوجي جو استعمال مذهبي معاملن ۾ به ٿيڻ لڳو آهي	9.
اسانکي اعتراف ڪرڻ گهرجي ته انساني حقن جي احترام ڪرڻ ۾ ۽ انساني حقن ڏيڻ ۾ مجموعي طور تي اسين سڀ ناڪام ويا آهيون	10.

B. Process on Tool

Data process on tool knows as information technology. The text data processed include words, Sentences, Paragraph and documents having multiple forms. In this research study based on sentences form processed data into tool and get appropriate output after processing input data. Output of data is in the form of numerical.

C. Evaluation of Results

Evaluation of results is based on SindhiNLP tool and data set used as Sindhi language those data taken from online Awami newspaper website. The results are based on seven tasks such as: (Lemma, Stem Suffix, Stem Affix, Stem, SPOS, UPOS and words). Those tasks performed on SindhiNLP tool. SindhiNLP tool is freely available for Sindhi Language task performed and gets appropriate results. Lemma word used in natural language processing to analysis of text morphologically, Stem suffix means to connect words at start word from text and derivational morphemes, Suffix means formation of words, Stem Affix is a spring morpheme that is connected after, or within a root or stem, Sindhi Parts-of-Speech from Sindhi language, Eight parts of speech of English language is called universal parts-of-speech, to measure the words from input data of Sindhi Sentences [10] & [17].

III. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental results based on Sindhi Language to perform tasks on tool. Those results are also given in table 2.

In Table 2 research study based on Sindhi language and processing on Sindhi NLP tool performed seven different tasks on tool. In first task Lemma value of Sindhi data was obtained and all the sentences have different values of the task (lemma). Maximum value (22) of lemma was found in sentence no. 10 and minimum value (8) of lemma was of sentence no. 3 and 5.

In 2nd task of Stem suffix value of Sindhi data was obtained and all the sentences have different values of the task (Stem suffix). Maximum value (12) of Stem suffix was found in sentence no. 10 and minimum value (3) of lemma was of sentence no. 1, 2 & 3.

In 3rd task of Stem affix value of Sindhi data was obtained and all the sentences have different values of the task (Stem suffix). Maximum value (1) of Stem affix was found in sentence no. 1, 2, 5 &10 and other sentences have no any value because tool did not find their stem affix values.

In 4th task of Stem value of Sindhi data was obtained and all the sentences have different values of the task (Stem).

Maximum value (22) of Stem was found in sentence no. 10 and minimum value (8) of Stem was of sentence no. 1 & 5.

5th task of SUPOS, 6th task of UPOS and 7th task Words having same value of Sindhi data was obtained and all the sentences have different value). Maximum value (22) was found in sentence no. 10 and minimum value (8) was of sentence no. 3 & 5.

TABLE II. EXPERIMENTAL RESULTS

Sindhi Experimental Results Performed on SindhiNLP tool							
S. No.	**L**	**SS**	**SA**	**S**	**SPOS**	**UPOS**	**W**
1.	13	3	1	13	13	13	13
2	17	3	1	17	17	17	17
3.	8	4	--	8	8	8	8
4.	21	10	--	21	21	21	21
5.	8	4	1	8	8	8	8
6.	17	5	--	17	17	17	17
7.	19	6	--	19	19	19	19
8.	19	6	--	19	19	19	19
9.	11	6	--	11	11	11	11
10.	22	12	1	22	22	22	22

IV. CONCLUSION AND FUTURE WORK

This research study based on Sindhi Language processing on tool and performed seven tasks that are: (Lemma, stem suffix, Stem Affix, Stem, SPOS, UPOS and words). Those tasks evaluate from Sindhi language sentences. Sindhi data set have used ten sentences and one hundred ninety-five words for research to find out root words from Sindhi data. In future work, take huge amount of Sindhi data set and performed natural language processing task by using different techniques.

ABBREVIATIONS

Artificial Intelligence	=	AI
Machine Learning	=	ML
Natural Language Processing	=	NLP
Sindhi Language	=	SL
Lemma	=	L
Stem Suffix	=	SS
Stem Affix	=	SA
Stem	=	S
Sindhi Parts-of-Speech	=	SPOS
Universal Parts-of-Speech	=	UPOS
Words	=	W

REFERENCES

[1] M. A. Dootio and A. I. Wagan, “Development of Sindhi text corpus,” *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.02.002.

[2] M. U. Rahman, “Sindhi Language Authority,” *Http://Www.Sindhila.Org*, 2010.

[3] M. Leghari and M. U. Rahman, “Towards Transliteration between Sindhi Scripts Using Roman Script,” *Linguist. Lit. Rev.*, vol. 1, no. 2, pp. 101–110, 2015, doi: 10.32350/llr.12.03.

[4] D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, and G. N. Mojai, “Issues and Challenges in Sindhi OCR,” *Sindh Univ. Res. J. (Science Ser.*, vol. 46, no. 2, pp. 143–152, 2014.

[5] M. Ali and A. Imdad, “Sentiment Summerization and Analysis of Sindhi Text,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 296–300, 2017, doi: 10.14569/ijacsa.2017.081038.

[6] Z. Rehman, W. Anwar, and U. I. Bajwa, “Challenges in Urdu Text Tokenization and Sentence Boundary Disambiguation,” *Work. South Southeast Asian NLP*, pp. 40–45, 2011.

[7] A. Ziani, N. Azizi, D. Zenakhra, S. Cheriguene, and M. Aldwairi, “Combining RSS-SVM with genetic algorithm for Arabic opinions analysis,” *Int. J. Intell. Syst. Technol. Appl.*, vol. 18, no. 1–2, pp. 152–178, 2019, doi: 10.1504/IJISTA.2019.097754.

[8] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, “Associating targets with SentiUnits: A step forward in sentiment analysis of Urdu text,” *Artif. Intell. Rev.*, vol. 41, no. 4, pp. 535–561, 2014, doi: 10.1007/s10462-012-9322-6.

[9] 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.

[10] I. N. Sodhar, A. H. Jalbani, A. H. Buller, E. Study, and S. Language, “AN EMPIRICAL AND STATISTICAL STUDY ON POS TAGGING OF SINDHI SOCIAL MEDIA TEXT,” *FOURRAFES*, vol. 241, no. 1, pp. 72–81, 2020.

[11] M. U. Rahman, “Towards Sindhi Corpus Construction,” *Linguist. Lit. Rev.*, vol. 1, no. 1, pp. 39–47, 2015, doi: 10.32350/llr/11/04.

[12] I. N. Sodhar, A. H. Jalbani, A. H. Buller, M. I. Channa, and D. N. Hakro, “Sentiment analysis of Romanized Sindhi text,” in *Journal of Intelligent and Fuzzy Systems*, 2020, doi: 10.3233/JIFS-179675.

[13] 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.

[14] T. Spilioti, “From transliteration to trans-scripting: Creativity and multilingual writing on the internet,” *Discourse, Context Media*, vol. 29, no. xxxx, pp. 1–10, 2019, doi: 10.1016/j.dcm.2019.03.001.

[15] A. Assiri, A. Emam, and H. Aldossari, “Arabic Sentiment Analysis: A Survey,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 12, pp. 75–85, 2015, doi: 10.14569/ijacsa.2015.061211.

[16] A. H. Suwaidi, T. R. Soomro, and K. Shaalan, “Sentiment analysis for Emiriti dialects in Twitter,” *Sindh Univ. Res. Journal-SURJ (Science Ser.*, vol. 48, no. 4, pp. 707–710, 2016.

[17] 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.