INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Exploration of Sindhi Corpus Through Statistical Analysis on the Basis of Reality

Irum Naz Sodhar[1]*, Suriani Sulaiman[2], Abdul Hafeez Buller[3]

1 Post-Doctoral Fellow, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University, Malaysia
2 Assistant Professor, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University, Malaysia
3 Post-Doctoral Fellow, Department of Civil Engineering, Kulliyyah (Faculty) of Engineering, International Islamic University, Malaysia

OPEN ACCESS

Received: 01-02-2023
Accepted: 23-03-2023
Published: 28-03-2023

Citation: Sodhar IN, Sulaiman S, Buller AH (2023) Exploration of Sindhi Corpus Through Statistical Analysis on the Basis of Reality. Indian Journal of Science and Technology 16(12): 924-931. https://doi.org/10.17485/IJST/v16i12.236

* **Corresponding author**.

iram10akber@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The Sindhi language is given more importance in Sindh's educational institutions than other regional languages, and the majority of the population uses it in today's mobile programs, letters, text messages and other text conversations. Research is needed to analyze the Sindhi corpus, as communication over computer systems and mobile phones is growing significantly. This research study focuses on the Sindhi alphabet and performs different tasks on the corpus. **Methods:** Data collection was conducted from available resources, and a corpus was created in Sindhi and English. Twenty patterns of letters are used, three dot alignments are used in the letters, and six symbols are used for making letters. After the collection, data was explored and analyzed with different tasks. **Findings**: The corpus of Sindhi text is being built due to its importance for language, linguistics and other developments in NLP. This research study focuses on statically analyzing the Sindhi-English corpus through reality basis, finding that there are two small words (۾ and ۽) and three biggest words (پاڪستان, انگلينڊ and ڳالهيون). The letter 'آ' is used as a single letter in Sindhi alphabets, with the minimum frequently occurring letter being consonant and the maximum frequently being ئ. vowel. **Novelty:** Text analysis is an important area in data mining and in other research, and this research study focuses on statically analyzing the Sindhi-English corpus through statically on reality basis. The author explores orthography and Sindhi composition of copra, and recommends that the Romanized languages data be used in Sindhi as well. Preprocessing is not easy due to lack of resources, and the character conversion model has generated two languages.

Keywords: Sindhi; Language exploration; Corpus; Statistical Analysis; pattern of letters; Text conversation

1 Introduction

The Indo-Aryan language family belongs to the wider family of Indo-European languages. The majority of these languages are spoken in Pakistan, Bangladesh, Nepal, and North and Central India. Indo-Aryan languages are also spoken in the Maldives and Sri Lanka, two neighboring island nations [1].

The major Indic languages have a rich history of writing and employ numerous scripts descended from the historic Brahmi script. These scripts all represent nearly the same set of phonemes and have good grapheme-to-phoneme correlation. Each script has a unique authorized range of code points in the Unicode standard due to the fact that the visual layout of the characters varies greatly between languages. Urdu is the most well-known of these, using a script borrowed from Arabic, while Kashmiri, Punjabi, and Sindhi are both Arabic- and Brahmi-derived [2]. Central Indian and North-Eastern Indian languages, which historically lacked a literary heritage, now use the Latin script or one of the many scripts descended from Brahmi [1].

Sindhi is a language of the Indo-Aryan family spoken mostly in Pakistan's Sindh Province, while in India, Gujrat, Rajasthan, and Maharashtra are the three most populous Sindhi-speaking regions [3]. It is the oldest spoken and written language in Pakistan and is given more weight than other regional languages in Sindh's educational institutions. The only regional tongue that was formally taught at educational institutions by the British was Sindhi. More than 30 million people live in Pakistan and 2.8 million people reside in India, with Sindhi as its official language and Urdu as the country's official tongue. The Sindhi language is the second most widely written language after Urdu [4].

The creation of a humanoid, regarded as the most intelligent machine, is the ultimate goal of artificial intelligence [5]. The interaction between people and computers needs to be carefully considered during this development process. In order to process communication languages across technological dimensions, it is desirable to develop language tools [6].

The success rate of natural language processing applications such as machine translation, search engines, social networking, word segmentation, proofreading, information retrieval, and natural language understanding is improved by spell checking, one of the most thoroughly explored NLP jobs [5].

Written communication requires scripts that include specific characters, literals, and rules. There are significant structural differences among these characters and rules. Indian and Pakistani languages have a large number of half letters, which contribute to their extensive vocabulary. Due to their extensive vocabulary, these languages are more challenging to learn than western languages like English [5].

This study examines the current state of Sindhi corpus construction and explores issues like corpus gathering, tokenization, and preprocessing. It also examines orthography and script in relation to corpus formation. Corpora, the plural form of corpus, are used to refer to extremely large text data sets that contain millions and billions of text records [7]. Different text corpora have been created in various languages of various countries languages [8]. The lack of resources for computational linguistics and research made the task of natural language processing particularly difficult [9].

Extensive, specific Sindhi computational linguistic research is crucial and vital for the development and maturity of the corpus. The Sindhi tag set must be designed prior to the POS of the corpus being tagged [10]. The statistical analysis of the Sindhi corpus is a subject that has to be thoroughly investigated [9].

This article addresses issues such as data mining and pre-processing when creating a Sindhi corpus [11]. The author explores the orthography and Sindhi composition of copra [12]. Preprocessing is not easy due to the minimum availability of resources for computational linguistics and exploration and the creation of different textual data in many languages in different countries [13]. The character conversion model has generated two languages. Analyzing language, the author recommended that the Romance language data be used in Sindhi as well.

2 Methodology

This research study focuses on Sindhi Language alphabet and perform different task on corpus as shown in Figure 1. Sindhi is divided into three groups (patterns, symbols and dots). In the first phase of the research study, which was data collection, which was collected from the AwamiAwaz newspaper [14], the corpus (Sindhi) was created as well as in English. In the Sindhi alphabet, twenty patterns of letters are used; three dot alignments (top, center and bottom) are used in the letters and six symbols are used to make the letters as shown in Figure 2. After the collection of corpus, data was explored statically and different tasks were performed (Total Letters in corpus, Repeat Letter Used in Corpus, Used Letters Particular in Corpus, Tokenization, Biggest Words and Smallest Words).

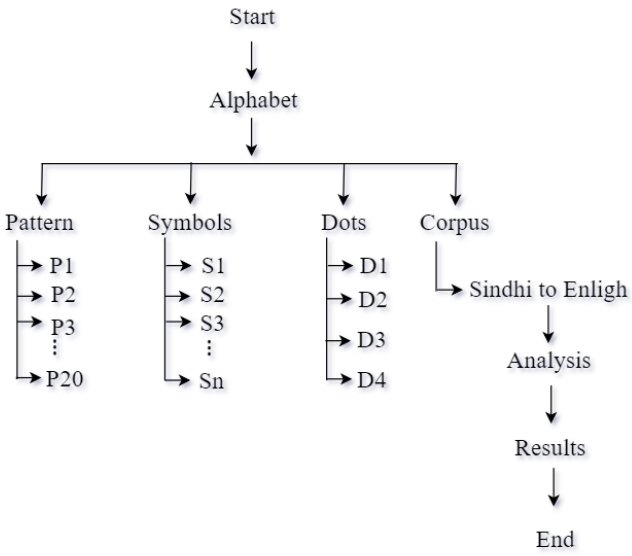

Fig 1. Research Methodology for exploration of corpus

2.1 Sindhi Alphabet

Sindhi was a difficult language to learn, as it consisted of 52 letters compared to 39 letters in Urdu, 32 letters in Persian, 28 letters in Arabic, and 26 letters in English. Its spelling and orthography were the same as Arabic, and it uses dots (single dot, double dot, triple dot, four dots) in Sindhi scripts. Several arrangements of dots were found, sometimes below, above, inside, and of the letters.

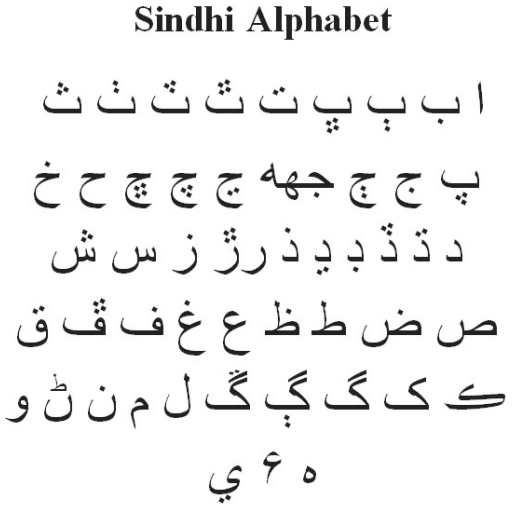

Fig 2. Sindhi Alphabet

Sindhi alphabet contains total fifty-two no. of letters to construct variety of words, sentences and documentations in Sindhi Language[15]. Sindhi alphabet contains number of dots with different position of letters such as: top of letter, central of letter and beneath of letter as shown in Figure 2. Sindhi letters have twenty patterns (P) of shapes which are described below Table 1.

Table 1. Patterns of Sindhi Alphabet

Pattern	Roman English	Sindhi
P-1	One letter of alphabet that is alif	ا
P-2	Nine letters of alphabet as ba, baa, bha, ta,tha, taa, tha, sah, pa	ب ٻ ڀ ت ٿ ٽ ٺ ث پ
P-3	Seven letters of alphabet as jum, jah, chah, cha, cha, ha, kha	ج ڄ ڃ چ ڇ ح خ
P-4	Six letters of alphabet as dal, dha, dah, daah, zal	د ڌ ڏ ڊ ڍ ذ
P-5	Three letters of alphabet as, ra, daah, zah	ر ڙ ز
P-6	Two letters of alphabet as, seen, sheen	س ش
P-7	Two letters of alphabet as, suwad, zuwad	ص ض
P-8	Two letters of alphabet as, tuwah , zuwah	ط ظ
P-9	Two letters of alphabet as, aeenh, gaeenh	ع غ
P-10	Two letters of alphabet as, feeh, faah	ف ڦ
P-11	One letter of alphabet as, kaaf	ق
P-12	One letter of alphabet as, kaf	ڪ
P-13	Four letters of alphabet as, kha, gah,gha, ghan	ک گ ڳ ڱ
P-14	One letter of alphabet as, laam	ل
P-15	One letter of alphabet as, meem	م
P-16	Two letter of alphabet as, naanh,noon	ن ڻ
P-17	One letter of alphabet as, wao	و
P-18	One letter of alphabet as, ha	ه
P-19	Comes one letter of alphabet as Hamza	ء
P-20	Comes one letter of alphabet as, yah	ي

2.2 Alignment of Dots

Basically, alignment of dots depends on three positions which are: Top of letter, Center of letter and beneath of letter as described in Table 2. In top of dots in letter comes twenty letters, Center of dots in five letters and beneath of dots in eight letters with different count of dots. Total number of dots used letter of Sindhi are: One dot letters with top position (ذ, خ, ز, ض, ظ, غ, ف, ن), secondly in beneath position letters with one dot (ب , ڊ) and thirdly with center position letters of one dot (ج). Two dot letters with top position (ت , ڌ, ق, ڱ) beneath position letters with two dots (ب, ڍ, ڳ , ي) and center position letters of two dots (ڃ, ڄ). Three dot letters with top position (ث, ڏ , ش, ٿ),, beneath position letters with three dots (ڀ) and center position letters of three dots (چ). Four dot letters with top position (ٺ, ڙ, ڦ), beneath position letters with four dots (ٻ) and center position letters of four dots (ڇ).

Table 2. Dots Alignment in Letters

Alignment of Dots		
1.	Top	ت ٿ ٽ ٺ ث خ ڌ ڏ ذ ڙ ز ش ض ظ غ ف ڦ ق ڱ ن
2.	Center	ج ڄ ڃ چ ڇ
3.	Beneath	ب ٻ ڀ پ ڊ ڍ ڳ ي

2.3 Symbols

Major symbol of the Sindhi language are: almandah, nanh, hamza ,zabar, zeer, peehu (آ, ڻ, ء, - , -, و) symbols used to in writing style to pronounce properly. Otherwise sense does not match which word has written.

2.4 Statistical Analysis of Sindhi-English Corpus

A key factor in starting an investigation is the availability of the appropriate dataset. The NLP method is also implemented in the dataset and the results are fully analyzed based on the dataset. The Sindhi training dataset is not yet available for research[16]. The major gaps in Sindhi language have been identified and resource availability is minimal. In below Table 3, Corpus of Sindhi-English contain Nineteen (19) words, sixty nine (69) letters with seventy two (72) dots two symbols and one full stop

(.) punctuation are used in the Sindhi sentence. In Sindhi text without, using of dots and symbol in writing style does not give any sense of understanding. In English corpus of Sindhi words contain seventeen (17) words with eighty (80) letters, four (04) dots and no any symbol. Only one punctuation full stop (.) is used, and the detailed description of Sindhu-English is shown in Table 4.

Table 3. Corpus of Sindhi-English

Language	Corpus
Sindhi [13]	پاڪستان ۽ انگلينڊ وچ ۾ ٽيندڙ سيريز جي شيڊول تي ٻنهي بورڊن وچ ۾ ڳالهيون آخري مرحلي ۾ آهن .
English	Talks between the two boards are in the final stages of the series between Pakistan and England.

Table 4. Exploration of Corpus of Sindhi-English

	Exploration of Corpus	
Sindhi	Total Letters in corpus	پ ا ڪ س ت ا ن ۽ ا ن گ ل ي ن ڊ و چ ۾ ٽ ي ن د ڙ س ي ر ي ز ج ي ش ي ڊ و ل ت ي ٻ ن ه ي ب و ر ڊ ن و چ ۾ گ ڀ ا ل ه ي و ن آ خ ر ي م ر ح ل ي ۾ آ ه ن
	Repeat Letter Used in Corpus	س ا ن ل ي ڊ و چ ۾ ر ه آ ت
	Used Letters Particular in Corpus	پ ڪ ۽ گ ٽ ڊ ڙ س ز ج ش ڊ ب ڀ گ خ ر ي م ح
	Tokenization	پاڪستان"-1، "۽"-2، "انگلينڊ"-3، "وچ"-4، "۾"-5، "ٽيندڙ"-6، "سيريز"-7، "جي"-8، "شيڊول"-9، "تي"-10، "ٻنهي"-11، "بورڊن"-12، "وچ"-13، "۾"-14، "ڳالهيون"-15"، "آخري"-16، "مرحلي"-17، "۾"-18، "آهن"-19،
	Biggest Words	"پاڪستان"-، "انگلينڊ"-، "ڳالهيون"-
	Smallest Words	"۽", "۾"
English	Total Letters in corpus	T, a, l, k, s, b, e, t, w, e, e, n, t, h, e, t,w, o, b, o, a, r, d, s, a, r, e, i, n, t, h, e, f, i, n, a, l, s, t, a, g, e, s,o, f, t, h, e, s, e, r, i,e, s,b, e, t, w, e, e, n,P, a, k, i,s, t, a, n, a, n, d, e, n, g, l, a, n, d.
	Repeat Letter Used in Corpus	t, a, i, k, s, b, e, w, n, h, o, r, d, i, f.
	Used Letters Particular in Corpus	P
	Tokenization	“Talks”-1,” between”-2,” the”-3,” two”-4,” boards”-5,”are”-6,” in”-7,” the”-8,” final”-9,” stages”-10,” of”-11,” the”-12,” series”-13 “between”-14,” Pakistan”-15,”and”-16,”England”-17.
	Biggest Words	Pakistan
	Smallest Words	In

According to and Khoso et al. [8] Dootio and Wagan [17] Sindhi has remained a language with limited resources and a poor socio-economic status. In order to develop various kinds of algorithms and NLP-based solutions for the resolution of Sindhi language difficulties, a scientific methodology is developed. An essential and fundamental component of the online Sindhi parser is the statistical analysis. It examines the tokens grammatically and morphologically. The number of morphological forms that match to Sindhi tokens is displayed by the morphological analyzer.

3 Results and Discussion

The corpus of Sindhi text is being built due to its importance for language, linguistics and other Natural Language Processing developments. Accessible resources on the internet are not sufficient to provide adequate data, but this is not an excuse to not work on it. That's why this research study focuses on statically analysis of Sindhi-English corpus through statically on reality basis. In Sindhi alphabet contain fifty two letters with three different positions of dot and two symbols. In English alphabet contains only twenty six letters with two (Capital and lower) forms. In capital letters of English alphabet no any dots but in lower form have two dots one in "i" and second "j".

Exploration of the corpus is based on statistically, and in below Figure 3 exploration of Sindhi-English corpus is presented. In the first phase of study discussed about Sindhi corpus performed six tasks based on the statistically such as: Total number of letters in corpus sixty nine with seventy two dots, total number of repeated letters thirteen, total particular letters twenty, tokenization of words nineteen, biggest words three and smallest words two were used. In second phase of study English corpus of Sindhi also performed six same tasks as Sindhi. Tasks are: total number of letters eighty in corpus, with four dots, total number of repeated letters fifteen, total particular letter one, tokenization of words eighteen, biggest word one and smallest word one.

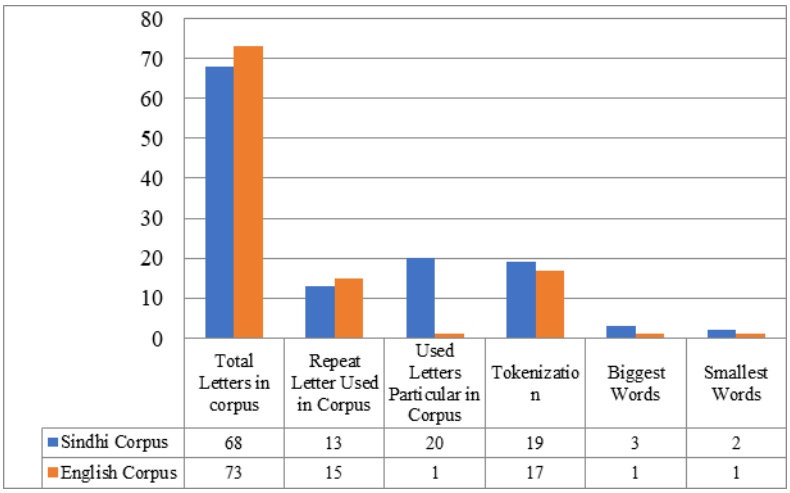

Fig 3. Exploration of Sindhi-English Corpus

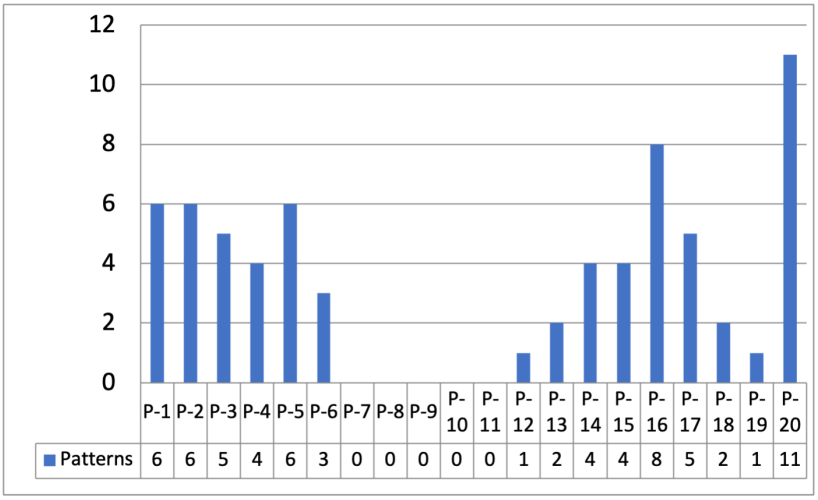

Fig 4. Patternsof Sindhi Corpus

After statically analysis of Sindhi corpus, it seems that there are two small words (۽ and ۾) are used, three biggest words (ڳالهيون and انگليند, پاڪستان) are used. These س ا ن ل ي ٻ و ڇ ۾ ر ه آ ت thirteen words are repeated words. Same way English one small word (in) and one biggest word (Pakistan) used in the corpus.

Patterns of Sindhi is plays an important role in analysis / exploration. There are twenty pat-terns are used in the Sindhi corpus as discussed in Table 1. In this research study fifteen pat-terns were explored in Sindhi corpus as shown in below Figure 4. P-20 is used eleven times in the Corpus, but P-7 to P-11 were not explored in Sindhi Corpus. Pattern

The letter 'آ' is used as a single letter in Sindhi alphabets, with the minimum frequently occurring letter being consonant were used ڳ and the maximum frequently being ئ. vowel is used in this research. Full Stop (.) is not important for information retrieval, search engines, and other text analysis processes. They are used to complete sentences and give a sense of understanding.

4 Conclusion

This research study has focused on statically analyzing the Sindhi-English corpus through statically on reality basis. The first phase of study discussed six tasks based on the statistically such as total number of letters in corpus sixty nine with seventy two dots, total number of repeated letters thirteen, total particular letters twenty, tokenization of words nineteen, biggest words three and smallest words two. Patterns of Sindhi play an important role in analysis and exploration, with twenty patterns used in the Sindhi corpus. Full stop (.) words are not important for information retrieval, search engines, and other analysis processes of the tax, but are used to complete sentences and give a sense of understanding. Pattern P-20 is used eleven times in the corpus, but patterns P-7 to P-11 were not explored in the Sindhi corpus. Statistical analysis of the text is an important area in data mining and research because it helps organizations extract useful data and information from text corpora. Future work research can lead to analyzing the Sindhi Corpus statistically and can provide a solution for those who worked on corpus tasks in different domain.

5 Acknowledgement

This research was supported by International Islamic University Malaysia to give a platform for the Post-Doctoral research. The authors are thankful to their colleagues, Dr. Suriani Sulaiman and Dr. Akhtar Hussain Jalbani, who provided expertise that greatly assisted the research. This research was not funded by any Institution/agency/company, but self-funded by the corresponding author and there exists no conflict of interest.

References

1) Kunchukuttan A, Bhattacharyya P. Utilizing language relatedness to improve machine translation: A case study on languages of the Indian subcontinent. 2020. Available from: https://doi.org/10.48550/arXiv.2003.08925.
2) Ashraf H. The ambivalent role of Urdu and English in multilingual Pakistan: a Bourdieusian study. *Language Policy*. 2022;22:1–24. Available from: https://doi.org/10.1007/s10993-022-09623-6.
3) Iyengar A, Parchani SL. Like Community, Like Language: Seventy-Five Years of Sindhi in Post-Partition India. *Journal of Sindhi Studies*. 2021;1(1):1–32. Available from: https://doi.org/10.1163/26670925-bja10002.
4) Nawaz A, Shaikh RA, Arain RH, Rajper S, Baber J, Baidani MM. Text Summarizer for Sindhi Language. .
5) Singh S, Singh S. Systematic review of spell-checkers for highly inflectional languages. *Artificial Intelligence Review*. 2020;53(6):4051–4092. Available from: https://doi.org/10.1007/s10462-019-09787-4.
6) Sodhar IN, Bhanbhro H, Amur ZH, Jalbani AH, Buller AH. Sindhi Language Processing on. *Online SindhiNLP Tool University of Sindh Journal of Information and Communication Technology vol*;4:4–7. Available from: https://doi.org/10.13140/RG.2.2.36489.47203.
7) Palh RB, Nawaz H, Shaikh ZA, Wagan AA. Design and Develop CMS for Sindhi E-News Papers. *Indian Journal of Science and Technology*. 2019;12(46):01–08. Available from: https://doi.org/10.17485/ijst/2019/v12i46/148128.
8) Sodhar IN, Sulaiman S, Buller AH, Sodhar AN. Aspect-Based Sentiment Analysis of Sindhi Newspaper Articles. . Available from: https://doi.org/10.22937/IJCSNS.2022.22.5.54.
9) Khoso FH, Memon MA, Nawaz H, Musavi SH. To build corpus of Sindhi language. . 2019. Available from: https://dialnet.unirioja.es/servlet/articulo?codigo=6933916.
10) Sodhar IN, Jalbani AH, Channa MI, Hakro DN. Romanized Sindhi Rules for Text Communication. *April 2021*. 2021;40(2):298–304. Available from: https://doi.org/10.22581/muet1982.2102.04.
11) Sodhar IN, Jalbani AH, Buller AH, Sodhar AN. Data mining security for big data. CRC Press. 2022. Available from: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003107286-5/data-mining-security-big-data-irum-naz-sodhar-akhtar-hussain-jalbani-abdul-hafeez-buller-anam-naz-sodhar.
12) Khan MA, Zaki S. Corpus Assisted Critical Discourse Analysis of Pakistan's Language Education Policy Documents: What are the Existing Language Ideologies? *SAGE Open*. 2022;12(3). Available from: https://doi.org/10.1177/21582440221121805.
13) Sodhar IN, Jalbani AH, Buller AH, Channa MI, Hakro DN. Sentiment analysis of Romanized Sindhi text. *Journal of Intelligent & Fuzzy Systems*. 2020;38(5):5877–5883. Available from: https://doi.org/10.3233/JIFS-179675.

14) Awamiawaz. 2022. Available from: https://awamiawaz.pk/900983/.
15) and. Govt. of Sindh . 2022. Available from: https://www.sindh.gov.pk/history.
16) Sodhar IN, Buller AH, Sodhar AN. Identification of Online Statistical Translation and Text Issues in Communication Technologies. *International Journal of Advanced Trends in Computer Science and Engineering*. 2021;10(2):446–453. Available from: https://doi.org/10.30534/ijatcse/2021/021022021.
17) Dootio MA, Wagan AI. Development of Sindhi text corpus. *Journal of King Saud University - Computer and Information Sciences*. 2021;33(4):468–475. Available from: https://doi.org/10.1016/j.jksuci.2019.02.002.