Data in Brief 19 (2018) 1504–1514

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Unicode-8 based linguistics data set of annotated Sindhi text

Mazhar Ali Dootio [a,b,*], Asim Imdad Wagan [c]

[a] Shaheed Zulifqar Ali Bhutto Institute of Science & Technology (SZABIST), Karachi, Sindh, Pakistan
[b] Benazir Bhutto Shaheed University Lyari, Karachi, Sindh, Pakistan
[c] Mohammad Ali Jinnah University, Karachi, Sindh, Pakistan

ARTICLE INFO

Article history:
Received 30 September 2017
Received in revised form
1 May 2018
Accepted 15 May 2018
Available online 22 May 2018

Keywords:
Sindhi
NLP
Computational linguistics
Morphology
Lexicon
Dataset

ABSTRACT

Sindhi Unicode-8 based linguistics data set is multi-class and multi-featured data set. It is developed to solve the natural languages processing (NLP) and linguistics problems of Sindhi language. The data set presents information on grammatical and morphological structure of Sindhi language text as well as sentiment polarity of Sindhi lexicons. Therefore, data set may be used for information retrieving, machine translation, lexicon analysis, language modeling analysis, grammatical and morphological analysis, Semantic and sentiment analysis.

Table 1 Specifications of Data set

Subject area	*Natural Languages Processing*
More specific subject area	*Tagging, syntactic, Sentiment and Morphology Analysis of Sindhi Text*
Type of data	*Textual*
How data was acquired	*Corpus is taken from Sindhi newspapers, blogs and social media sites like* • http://sindhsalamat.com/ • http://awamiawaz.com/ • https://thefocus.wordpress.com/

* Corresponding author at: Shaheed Zulifqar Ali Bhutto Institute of Science & Technology (SZABIST), Karachi, Sindh, Pakistan.
E-mail addresses: mazharaliabro@gmail.com, mazharaliabro@bbsul.edu.pk (M.A. Dootio), aiwagan@gmail.com (A.I. Wagan).

https://doi.org/10.1016/j.dib.2018.05.062

	• http://www.thekawish.com/beta/ *The corpus is processed for NLP operations such as sentiment and morphological analysis, UPOS and SPOS tagging, lemma and stemming identification.*
Data format	*Data is in csv format*
Experimental factors	*Tagging, syntactic parsing, sentiment analysis, morphological analysis, lemmatization, stemming and lexicon analysis.*
Experimental features	*Unigram based analysis, token analysis, Tagging with UPOS and SPOS, Sentiment classification and morphological classification and analysis, Lemma and stemming identification*
Data source location	*Karachi, Sindh, Pakistan*
Data accessibility	*Data set may be downloaded from* http://www.sindhinlp.com/ *and github*

Value of the data

- Data set is developed on basis of acquired results of Sindhi online natural languages processing (NLP) tool for parsing, tagging, morphological and sentiment analysis, stemming and lemmatization of Sindhi text.
- Data set is valuable to comprehend the grammatical, sentimental, syntactic and morphological structure of Sindhi text.
- Dataset is significant source for machine learning and NLP analysis for information retrieving, language modeling, machine translations, sentiment analysis and computational linguistics operations.

1. Data

More research work has been done on English language [1] thus, lot of NLP resources are available for English language, which are not suitable for other languages such as Sindhi language. Right hand written languages are also important for NLP applications, machine and deep learning processes. Sindhi language is right hand written language and using Arabic-Persian writing style [2]. A good number of websites, blogs and social media pages are available on world wide web (www), thus, there is very good number of data available for computational linguistics, NLP, machine translations, information retrieving and machine learning processing. Polarity, UPOS annotation, SPOS annotation, Lemma and Stemming process for Sindhi text. Sindhi NLP tools are used to annotate Sindhi corpus for various purposes like tagging, sentiment analysis, lemma and stemming identification and etc. Fig. 1 shows annotation process for Sindhi text وڏن جو احترام ڪرڻ هڪ سٿو عمل آهي ۽ اسان تي فرض به آهي (Waddan jo ahtaraam karann hik sutho amal aahay aen asaan te farz be aahay).

Fig. 2 shows the morphological analysis of Sindhi text ۽ اسان تي فرض به آهي وڏن جو احترام ڪرڻ هڪ سٿو عمل آهي (Waddan jo ahtaraam karann hik sutho amal aahay aen asaan te farz be aahay).

Fig. 3 shows the sentiment analysis [3] of Sindhi text ۽ اسان تي فرض به آهيبه آهي وڏن جو احترام ڪرڻ هڪ سٿو عمل آهي (Waddan jo ahtaraam karann hik sutho amal aahay aen asaan te farz be aahay)

The dataset is consisted of 19 attributes and 6841 records. Target classes of dataset are categorical therefore, it may be good for supervised analysis. Table 1 shows the statistical analysis of Class attributes of dataset.

Sindhi Corpus Annotation

Lemma	Stem Suffix	Stem Affix	Stem	SPOS	UPOS	Word	Word ID
ونڻ	ان		وڻ	صفت	ADJ	ونڻ	305
جو			جو	حرف جر	ADP	جو	70
احترام			احترام	اسم	NOUN	احترام	1180
ڪر	اڻ		ڪر	فعل	VERB	ڪرڻ	159
هڪ			هڪ	صفت عددي	NUM	هڪ	82
سنو	او		سُٹ	صفت	ADJ	سنو	32
عمل			عمل	اسم	NOUN	عمل	307
آهي	اي		آڌ	فعل معاون	AUX	آهي	2
۽			۽	حرف جملو	CONJ	۽	83
اسان		ا	سان	ضمير	PRON	اسان	69
تي	ي		ت	حرف جر	ADP	تي	13
فرض			فرض	اسم	NOUN	فرض	308
به			به	حرف جر	ADP	به	309
آهي	اي		آڌ	فعل معاون	AUX	آهي	2

Fig. 1. Annotation process for Sindhi corpus.

Morphological Words	Total number in text	Percentage
Simple Words	11	78.57
Complex Words	2	14.29
Compound Words	2	14.29
Reduplicated Words	0	0.00

Fig. 2. Morphological analysis of Sindhi corpus.

2. Experimental design, materials and methods

Sindhi corpus documents are processed for annotation and sentiment analysis in Sindhi NLP tool separately. The results of annotation and sentiment analysis are accumulated to develop dataset. Unigram model is used to find probability of each lexicon in corpus. Dataset is processed for normalization and statistical analysis. There is no missing value found in the dataset. Brief introduction of attributes is given below:

1. **UPOS**: Universal Part of speech tag set [4,5] is used to annotate the Sindhi tokens. UPOS is class attribute, which is consisted of 18 categories. Sindhi tokens are tagged properly with UPOS tag set.

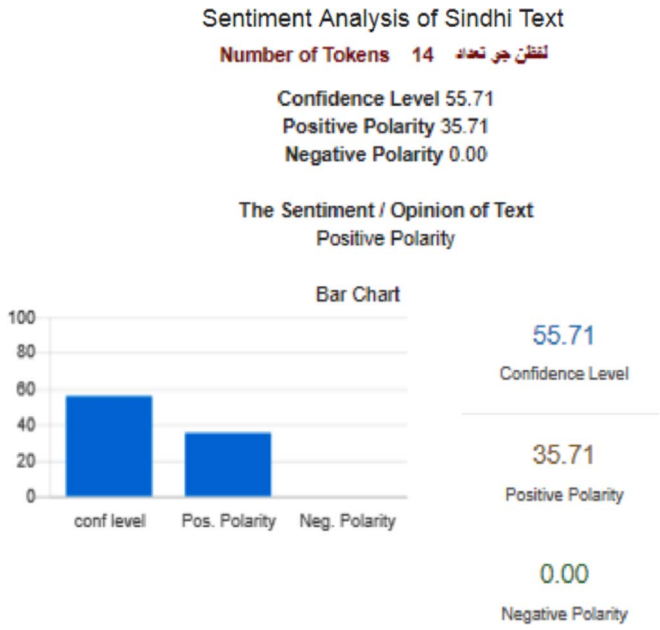

Fig. 3. Sentiment analysis of Sindhi corpus.

Table 1
Statistics of Sindhi annotated dataset.

Statistics	*UPOS*	*SPOS*	*Gender*	*Number*	*Polarity*	*Morphology*	*Lemma*	*Diacritic*	*Infinitive*
Count	6617	6528	6841	6841	6841	6841	6841	6841	6841
Mean	5.52	5.54	0.33	0.98	2.27	1.23	0.711	0.02	0.02
Std	4.36	4.37	0.65	0.45	1.08	0.63	0.45	0.15	0.15

Table 2
UPOS tagging to Sindhi text.

UPOS Tag	Complete Name	Sindhi token	English meaning
NOUN	Noun	انب	Mango
ADJ	Adjective	سڻو	Good
NOUN	Noun	ميوو	Fruit
AUX	Auxiliary Verb	آهي	Is
PERIOD	Period	.	Full stop

For example, Sindhi sentence انب سڻو ميوو آهي. (Mango is good fruit.) may be tagged with UPOS and Sindhi part of speech (SPOS) as shown in Table 2.

The frequency of UPOS tags is dissimilar from each other in the dataset, which shows the divergence of Sindhi lexicons. Fig. 1 shows the frequency of UPOS tag set, annotated to Sindhi tokens. Fig. 4 presents the high number of Nouns and low number of Subordinating conjunction.

2. **SPOS**: Sindhi Part of Speech (SPOS) tag set is indigenous Sindhi language tag set. SPOS is class attribute of the dataset and consisted of 17 categories. There is little difference between UPOS tag

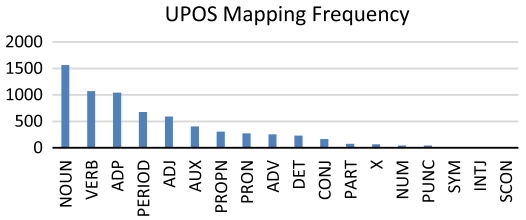

Fig. 4. Frequency of UPOS tagged to Sindhi tokens.

Table 3
SPOS tagging to Sindhi text.

SPOS Tag	English Name	Sindhi token	English meaning
اسم	Noun	انب	Mango
صفت	Adjective	سٹو	Good
اسم	Noun	ميوو	Fruit
فعل معاون	Auxiliary Verb	آهي	Is
پورو دم	Period	.	Full stop

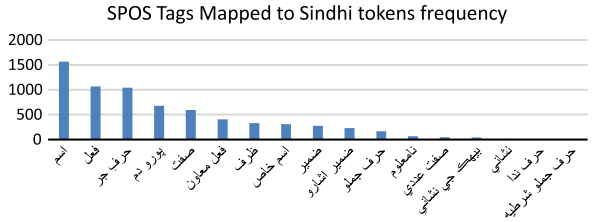

Fig. 5. Frequency of SPOS tagged to Sindhi tokens.

PART and SPOS tag Adverb and Preposition. PART POS annotates Sindhi negation and possessive lexicons, whereas, SPOS adverb POS annotates Sindhi negation lexicons and Preposition POS annotates possessive markers available in Sindhi language, therefore, Sindhi adverb and preposition are used in place of PART POS of UPOS tag set. Sindhi treebank is novel contribution to NLP because it is not used properly for the purpose of computational linguistics operations.

The frequency of SPOS tags is different than the frequency of UPOS tags because of difference of UPOS tag PART and SPOS tag Adverb. Table 3 shows the annotation process of SPOS to Sindhi text.

Fig. 5 shows the frequency of Sindhi POS tag set which annotated to Sindhi text document **آهي. انب سٹو ميوو** (Mango is good fruit.). There is high number of Nouns (**اسم**) and low number of frequency of Subordination conjunction (**حرف جملو شرطيه**) found in the dataset. The difference of frequency is obvious in UPOS Adverb and SPOS Adverb (**ظرف**).

3. **Gender**: According to Sindhi Grammar [6,7], there are two types of Gender. One is masculine called **جنس مذڪر (Jins Muzkar)** in Sindhi and second is feminine called **جنس مونث (Jins Moans)** in Sindhi. Noun, adjective and diacritic change the position of gender from masculine to feminine and vice versa. This attribute is class attribute and presents the lexicons with its proper gender. Table 4 shows the examples of Sindhi lexicons and their gender.

Table 4
Mapping of genders to Sindhi lexicons.

Sindhi lexicons	English Meaning	Gender	English meaning
ميز	Table	مونث	Feminine
موسم	Season	مونث	Feminine
انب	Mango	مذكر	Masculine
چوڪرو	Boy	مذكر	Masculine
چوڪري	Girl	مونث	Feminine
ڊاڪٽر	Doctor	مذكر	Masculine
ڊاڪٽرياڻي	Doctor	مونث	Feminine

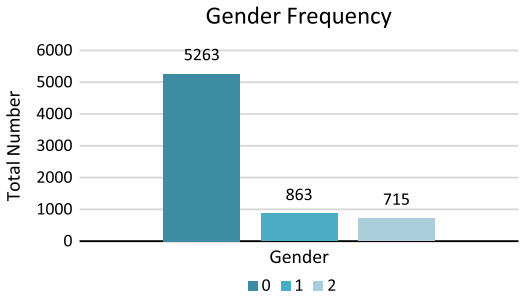

Fig. 6. Frequency of gender types.

This dataset shows Masculine gender with digit 1, feminine gender with digit 2 and digit 0 shows no gender, which is used for periods, punctuations and symbols. Fig. 6 shows the number of frequencies of gender types.

4. **Number:** Number is of two types in Sindhi Grammar [6,7]. One is singular (عدد واحد) and second is plural (عدد جمع).. This feature of Sindhi dataset shows the singular or plural status of lexicon. Diacritics and extensions of words such as يا ، آ، يون، اون make the plural number of nouns, pronouns and adjectives. Table 5 shows the status of singular and plural numbers of Sindhi text.

Singular number of noun, pronoun and adjective is shown with digit 1 and plural number of noun, pronoun and adjective is shown with digit 2 whereas, o shows no number, which is used for periods, punctuations and symbols. Fig. 7 shows the total number of frequencies of singular and plural number types.

5. **Polarity**: Polarity is class feature of the dataset, which is comprised of three categories: Positive, Negative and Neutral. All these types of polarity show the sentiment of lexicons. For example Sindhi lexicon سڻو (Sutho) shows positive polarity, خراب (Kharaab) shows negative polarity and ٺيڪ (Theek) shows neutral polarity. For example, Sindhi sentence انب مڻو ميوو آهي **(Anb mitho mevo aahay)** shows positive polarity because adjective مڻو (Mitho) shows positive sentiment. Digit 1 shows positive polarity, digit 2 shows negative polarity, digit 3 shows neutral polarity whereas, 0 shows no polarity which is used for periods, punctuations and symbols. Fig. 8 shows the total number of frequencies of Polarity types.

Table 5
Mapping of singular and plural numbers to Sindhi lexicons.

Sindhi lexicons	English Meaning	Number	English meaning
ميز	Table	عدد واحد	Singular noun
ميزون	Tables	عدد جمع	Plural noun
انب	Mango	عدد واحد	Singular noun
انبَ	Mangoes	عدد جمع	Plural noun
چوڪرو	Boy	عدد واحد	Singular noun
چوڪرا	Boys	عدد جمع	Plural noun
چوڪر	Girl	عدد واحد	Singular noun
چوڪريون	Girls	عدد جمع	Plural noun
سڻي	Good	عدد واحد	Singular adjective
سڻيون	Good	عدد جمع	Plural adjective

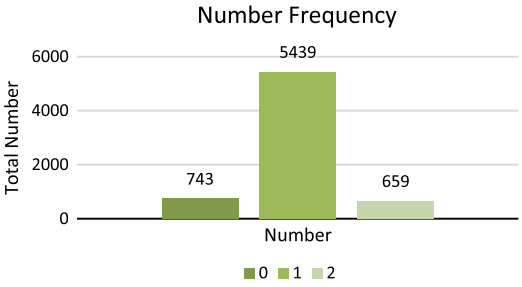

Fig. 7. Total frequency of number types.

6. **Morphology:** This feature of dataset presents the morphological form of Sindhi lexicon. The attribute shows both forms of morphology, which are free form and bound form. For example, Sindhi lexicon **يقين** (trust) is a free form and **سدائين** (ever) is a bound form. Bound form may be called secondary form which is divided into three categories [7], Complex, Compound and Reduplicated. Free form is shown with digit 1 and bound or secondary form is shows with digit 2 whereas, digit 0 is used for punctuations, symbols and periods. Fig. 9 shows the total number of frequencies of Morphology forms.

   a. **Complex words (مرتب لفظ):** Addition of affix or suffix to free form lexicon makes it the complex word of bound form. For example, Sindhi lexicon **ڃاڻ** (jjaann) may be complex word by adding affix ا (aa) which changes the word **ڃاڻ** to **اَڃاڻ**. Digit 1 shows status of lexicon as complex word and digit 0 shows no complex word.
   b. **Compound words (مرڪب لفظ):** It is combination of two free forms, which shows single meaning. For example, Sindhi compound word **گهرڌڻي** (House Owner) is combination of two free forms, **گهر** (ghar) means House and **ڌڻي** (Dhanni) which means owner. This attribute shows the feature of Sindhi lexicon that either it is compound or not. Compound lexicons show with digit 1 and non-compound lexicons are shows with digit 0.
   c. **Reduplicated words دهرايل يا ٻٽا لفظ:** There is little number of reduplicated words in this data set. Reduplicated words are basically compound words but the structure and presentation of these words are changed from compound words. For example, Sindhi words **اچڻ هنڌ هنڌ ، وک وک، اچ وچ ، راندروند** and etc. are reduplicated words of bound form.

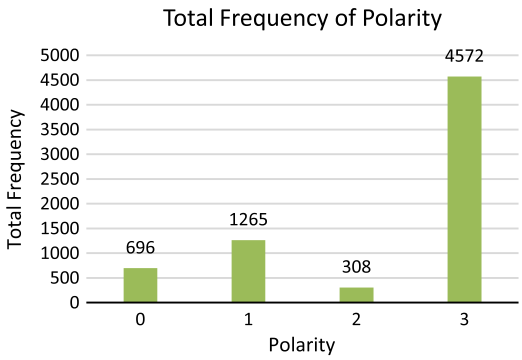

Fig. 8. Total frequency of number types.

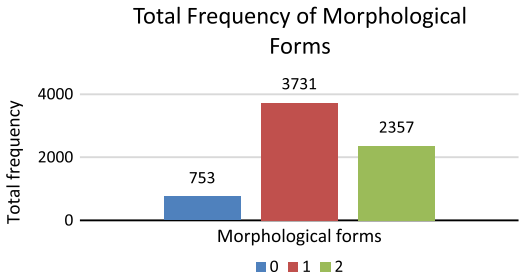

Fig. 9. Total frequency of morphology forms.

Reduplicated words are the feature of Sindhi text, which make Sindhi text complex to process for NLP and computational linguistics operations. This dataset presents reduplicated words with digit 1 and no reduplicated words with digit 0.

Fig. 10 shows the total number of frequencies of Morphology bound form types.

7. **Lemmatization**: Lemmatization is process of identifying the original lexicon by reducing affix or suffix which holds grammatical and morphological structure whereas, stemming reduce the inflections, diacritics, affixes and suffixes of lexicon and derive root word. For example, Sindhi word اچڻ (To come) is complex word, therefore, the lemma of this word is اچ (come). Word هو (Achu) shows complete meaning with proper grammar and morphological structure. Digit 1 shows lexicon as lemma and digit 0 shows no lemma. Fig. 11 shows the total number of frequencies of Lemma in the dataset.

8. **Diacritic**: Diacritic changes meaning of lexicon by attaching glyph to word or letter. For example, Sindhi lexicon هُو (was) is changed to another Sindhi lexicon ھُو (he) by attaching glyph ُ to Sindhi letter ھ (ha). The diacritic changes the meaning and grammatical structure of Sindhi lexicon هو (was). The first Sindhi lexicon هو (was) is verb which shows action happened in past and second Sindhi lexicon هُو (he) is determiner. This attribute of dataset shows the diacritic feature of Sindhi lexicon with digit 1 and shows no diacritic feature with digit 0. Fig. 12 shows the total number of frequencies of diacritic words.

9. **Infinitive**: Sindhi linguists give importance to infinitive verbs called in Sindhi مصدر (Massdar). Sindhi infinitive verbs are generated by attaching suffix to stemming or lemma words. Table 6 shows the example of Sindhi infinite verbs.

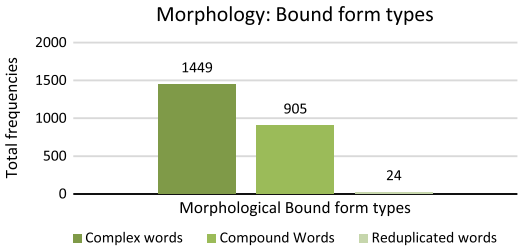

Fig. 10. Total frequency of morphology bound form types.

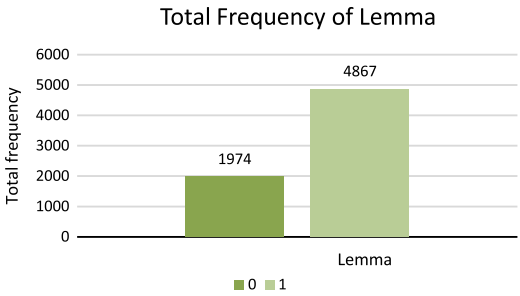

Fig. 11. Total frequency of lemma.

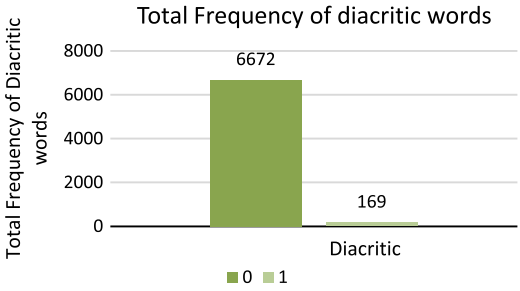

Fig. 12. Total frequency of diacritic words.

Fig. 13 shows the total number of frequencies of infinitive words available in Sindhi annotated dataset. 0 shows non-infinitive words and 1 shows infinitive words

10. **Sindhi Token:** This attribute is significant attribute of the dataset because all features and target classes are related to this attribute. Tokens are taken from Sindhi corpus. Sindhi tokens are of following types.
    a. Sindhi lexicons.
    b. Punctuations, Symbols and Periods.

Uni-gram probability is measured by applying statistical language model to display the impact of Sindhi tokens in the dataset.

Table 6
Example infinitive verbs of Sindhi text.

Lemma word	Infinitive Verb	English Meaning
اچ (Achu)	اچڻ (Achann)	To come
پڑھ (Parh)	پڑھڻ (Parhann)	To read
لک (Likh)	لکڻ (Likhann)	To write
کل (Khil)	کلڻ (Khilann)	To laugh
سمجھ (Samjh)	سمجھڻ (Samjhann)	To understand

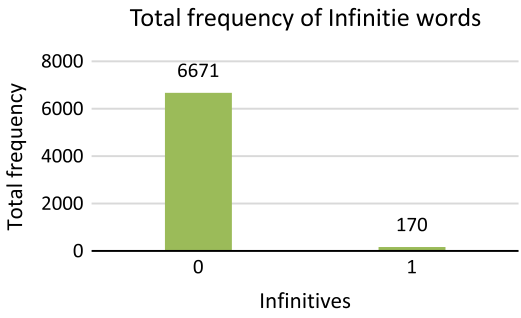

Fig. 13. Total frequency of diacritic words.

Acknowledgements

I confirm that our research paper “Syntactic parsing and supervised learning of Sindhi text” published in Journal of King Saud University -Computer and Information Sciences in October 2017, has used some portion of this dataset for the purpose of supervised analysis. However, dataset is modified and updated to publish in DIB journal. Some classes and feature attributes are added and updated as well as some attributes are removed and modified. Therefore, this dataset is updated version.

This research paper is produced from my doctor study on “Sentiment Analysis for Sindhi text” which is continued at SZABIST Karachi Sindh Pakistan. I am grateful to Dr. Hussnain Mansoor Ali Khan, program coordinator and Dr. Imran Amin Head of department of Computer science, SZABIST Karachi Sindh Pakistan for their support and provision of resources.

I confirm that we have not got any financial assistance or fund for this research study from SZABIST or any institution or organizations.

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.05.062.

References

[1] R. Tsarfaty, D. Seddah, S. Kübler, J. Nivre, Parsing morphologically rich languages: introduction to the special isue, Comput. Linguist. 39 (1) (2013) 15–22.
[2] M. Ali, A.I. Wagan, Syntactic parsing and supervised analysis of Sindhi text, J. King Saud. Univ. - Comput. Inf. Sci. (2017).

[3] M. Ali, A.I. Wagan, Sentiment summerization and analysis of Sindhi text, Int. J. Adv. Comput. Sci. Appl. 8 (10) (2017) 296–300.
[4] S. Petrov, D. Das, R. McDonald, A Universal Part-of-Speech Tagset, 2011.
[5] Y. Zhang, R. Reichart, R. Barzilay , and A. Globerson, Learning to map into a universal POS tagset, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 1368–1378, 2012.
[6] Mirza Kaleech Bag, Sindhi vya karan, Fourth. Jamshoro: Sindhi Adabi Board Jamshoro Sindh Pakistan, 2015.
[7] Dr. Ghulam Ali Alana, Sindhi Boli jo Tashreehi grammar (A descriptive grammar of Sindhi language), First. Jamshoro: Sindhi Lang. Auth., Hyderabad, Sindh Pak. (2010).