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.