Journal of King Saud University – Computer and Information Sciences 33 (2021) 468–475

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Development of Sindhi text corpus

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

[a] Shaheed Zulifqar Ali Bhutto Institute of Science and Technology 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 3 November 2018
Revised 22 January 2019
Accepted 3 February 2019
Available online 11 February 2019

Keywords:
Text corpus
NLP
Sindhi
DTM
TF-IDF

ABSTRACT

Sindhi language is a rich language with plenty of literary and general texts. There are number of books, newspapers, magazines and internet material available to develop Sindhi text corpus but yet proper and useful text corpus could not be developed and presented online for research, language features analysis, linguistics analysis and information retrieval systems. The lack of resources for research on computational linguistics and NLP applications for Sindhi language are challenging tasks at this stage. However, we have developed Sindhi text corpora in order to provide text resources to computational linguists, Natural Languages process (NLP) experts and researchers. Online books, newspapers, magazines, blogs and social websites are utilized to build Sindhi text corpus. Sindhi sentiment based text corpus is developed and analyzed with Document Term Matrix and TF-IDF models using 2-gram technique of n-gram model. The corpus may be useful for research on language variation analysis, sentiment analysis, aspect based sentiment analysis, semantic analysis, machine translation, information retrieval, Word2Vec, topic modeling and cluster analysis.

1. Introduction

Language is derived from the set of symbols which may be used for written or spoken types of language. It is basic and significant resource of human society for communication and business deal. People share their thoughts, values and resources through their languages. According to Dr. Alana (2010), language is an evolutionary process and problem of human beings like other problems of humans. The process and problem of language run along with humans from their birth to end of life. The process of linguistics makes it more perfect and scientific. Linguistics is not just a study of grammar but a scientific study of a language, which enables one to obtain features of language, enhance capability of communication process and provides one with deep learning of several aspects and functions of language. Computational linguistics solves and focuses the human languages problems, theoretical issues, cognitive issues and practical issues of languages using different computational applications, therefore, computational linguistic process works on theoretical and applied components of a language. Due to development of natural languages process and computational linguistics, different communities and nations of the world understand the languages of each other easily. Hence; latest development in machine translation, information retrieval, text analysis, text segmentation, syntactic parsing and etc. are done due to computational linguistics and natural languages process developments and research works. These developments and research works are motivations to technological less resourced languages of the world, such as Sindhi language. Now a day, research work is in progress on Sindhi language (Dootio and Wagan, 2019; Ali and Wagan, 2017; Jumani et al., 2018; Shah et al., 2018; Ali and Wagan, 2019; Dootio and Wagan, 2018) to evaluate and analyze its linguistics problems. This research study is part of current motivation for working on development of Sindhi text corpus.

1.1. Text corpus

Corpus of any language may be significant and important part of a language, on which research is done or going on, because practical investigations of language may be done using the corpus of that language. A corpus is body or a huge group of written text, which is developed for the purpose of linguistics analysis

* Corresponding author.
E-mail addresses: mazharaliabro@gmail.com, mazharaliabro@bbsul.edu.pk (M.A. Dootio).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.02.002

(Kennedy, 2014). Therefore, it is significant data, which provides lexicographers, grammarians and other interested personals with very good explanation of a language. The analysis of corpus, delivers information of morphology, syntactic parsing, lexicons structure, semantic, pragmatic and other linguistics components. The linguistic corpus may be written or spoken and open or close. The close corpus is specific and bound while open corpus is not bound and specific to a topic. However, the corpus may be used for the purpose of information retrieval, machine translation, pattern recognition, speech recognition, text to speech and speech to text recognition and synthesis, feature extraction and analysis, vectorization, word to vector analysis, dictionary development, thesaurus and WordNet development, word tokenization, text tagging, text parsing, morphological analysis, machine learning process, classification, cluster analysis and etc.

There is no significant research work found on Sindhi text corpus, however, less research work has been done on information retrieval, syntactic parsing, sentiment analysis and machine translation for this language. This research study has developed Sindhi text corpus and extracted its features and variations. DTM and TF-IDF models are used to find out the terms and features of Sindhi language for information retrieval and feature based sentiment analysis.

2. Related work

Multilingual countries are good to understand and speak multi-languages, while it is difficult for computers to understand the variety of natural languages. The unicode solves the problems of languages to make them familiar to computer systems. The proper corpus development needs proper tools and techniques. Cristina Bosco discusses the development process of text corpus that it is developed in three steps, which are collection, annotation and analysis. The annotated corpora are suitable for the classification of sentiments (Bosco et al., 2013). S Sharoff defines four steps for the corpus development and analysis, which are data collection, basic corpus cleaning, linguistic processing, and corpus evaluation (Schäfer and Bildhauer, 2013). The corpus may be topic-wise and polarity-wise. Amitava Das and et.al discuss the topic-wise or polarity-wise corpus that it extracts the related information from document (Das et al., 2012). SS Agrawal and et.al conducted a research study on multilingual corpus to know the complexity and distinction of languages. They explain the results that Nepali language is more complex than the Hindi and Punjabi languages (Agrawal et al., 2014). Baseer et al. (2016) conducted their research study on Urdu script corpus development and analysis therefore, they train the machine for cluster analysis using K-means machine learning method and get good results.

These surveys show the importance of corpus in the language model and development. Sindhi is morphological and grammatical rich and complex language, therefore, to develop a text corpus and analyze its contents are, basically, to provide solution of the computational linguistics problems of Sindhi language. The structure of right hand written languages is same at some levels but there is difference of sounds, lexicon structures, morphological structures and grammar, hence; research work and some linguistics tools may be beneficial for each other at some levels but not useful completely. Alana (2010) marks Sindhi language morphologically rich and complex language as it is using all forms of morphology including reduplicated words and compound verbs in its text. Therefore, this study has developed DTM and TF-IDF models using 2-gram model to identify the complex and compound words from Sindhi text corpus. Motlani (2016) considers Sindhi language technological resource poor language as there is very little work has been done on the technological development of Sindhi language. Nonetheless, a little work has been done on Sindhi language in view of computational linguistics. Rahman, Mutee. worked on building of Sindhi corpus and describes the basic requirements for its development. In that research study basic ideas for Sindhi corpus development are discussed (Rahman, 2010). Unavailability of the proper Sindhi text corpus with Arabic-Persia script and significant analysis model to analyze the Arabic-Persia script based Sindhi text corpus is a great attraction to work on Sindhi text corpus development and its analysis. This study has designed its own model to build and analyze the Sindhi text corpus for language variation and sentiment analysis.

3. Material and methods

The text corpus development is done using text corpus building process techniques. The texts are collected from online Sindhi books, newspapers, websites and blogs. Sindhi text corpus is processed for tokenization and universal part of speech (UPOS) tagging, lemmatization, stemming, sentiment analysis, aspect based sentiment analysis, morphological analysis. Fig. 1 describes the process and flow of development of Sindhi text corpus including process techniques and sources of data collection. Process starts from problem understanding and ends at text corpus development.

3.1. Sindhi text corpus development

It is fact that the accessible resources on internet are not providing enormous amount of Sindhi text data, however, they are not insufficient to give excuse of not working on Sindhi text corpus construction and analysis. Viewing the importance of Sindhi text corpus for language, linguistics and other NLP developments, the Sindhi text corpus is built. The following steps and standards are taken to develop a Sindhi sentiment based text corpus for sentiment analysis, aspect based sentiment analysis, language variation and other future research.

3.1.1. Representativeness

Construction of text corpus is process of selecting the language, type of text, target population, number of samples or texts, and length of the sample. Therefore, corpus representativeness shows the structure of text corpus, population of text corpus, range of linguistic features which are distributed in text corpus population and number of words in text corpus. Sindhi text corpora is built

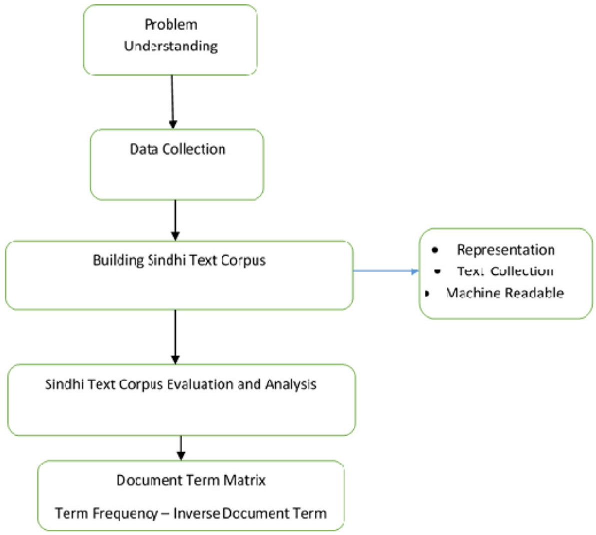

Problem Understanding
Data Collection
Building Sindhi Text Corpus
Representation
Text Collection
Machine Readable
Sindhi Text Corpus Evaluation and Analysis
Document Term Matrix
Term Frequency – Inverse Document Term

Fig. 1. Development process of Sindhi text corpus.

on basis of Sindhi story books, Socio-political analytic books, science books, fiction books, history books, linguistics books, literature books, newspaper articles, poetry, travelogues and blogs. Target population of Sindhi text corpus is patent and divided into several documents. Each text corpus document varies from other documents having dissimilar number of tokens. The research study has developed Sindhi text corpora to conduct more research on Sindhi language variations, feature distribution, syntactic analysis, POS tagging, stemming, lemmatization, semantic analysis, language modeling, machine translations, information retrieval and sentiment analysis. Fig. 2 shows the detail of topic-wise Sindhi text corpora.

However, the discussing and analyzing Sindhi text corpus is sentiment based text corpus, which describes the documents with positive and negative polarities. There are 11864 documents of positive polarity and 3924 documents are of negative polarity. Hence; the total number of documents in the text corpus is 15788. Overall, there are 23728 words which are showing positive polarity and 7848 words which are showing negative polarity in the text corpus. The corpus presents seven features of two electronic devices and the polarity of each feature opinion-wise. The opinion based polarity may be positive, negative or neutral. The electronic devices are Laptops and mobile cell phones. The features of these two devices are battery, mic, speakers, memory, camera, display screen and price.

3.1.2. Collection of text for Sindhi text Corpus development

Collection of proper and suitable text is significant process for the construction of a text corpus. To develop Sindhi text corpus, texts are collected through online resources. The collected texts are processed for construction of Sindhi text corpus. Each text corpus has finite size of documents. The size of text corpus depends on the type of documents, whereas, each document size varies from other document size.

Sindhi text corpus for sentiment analysis is part of Sindhi text corpora and it is developed by collecting textual data from social media sites such as Facebook and twitter, online newspapers, blogs, products websites, Sindhi websites and google forms. The reviews on products are available on social media in unstructured patterns, therefore, these reviews may be used for feature based or aspect based sentiment analysis (Krishna et al., 2018; Negi and Buitelaar, 2017). This text corpus provides current research study with products information including their aspects and polarity. The corpus presents the sentiments and opinions of users for laptops and mobile cell phones, therefore, it is closed corpus. The corpus documents present the opinions and sentiments of dissimilar users for both electronic products. Different types of mobile cell phones and computer laptop products are included in the Sindhi text corpus for aspect based sentiment analysis. The example of Sindhi text document (Fig. 3) is presented below to show the style and structure of Sindhi text corpus documents.

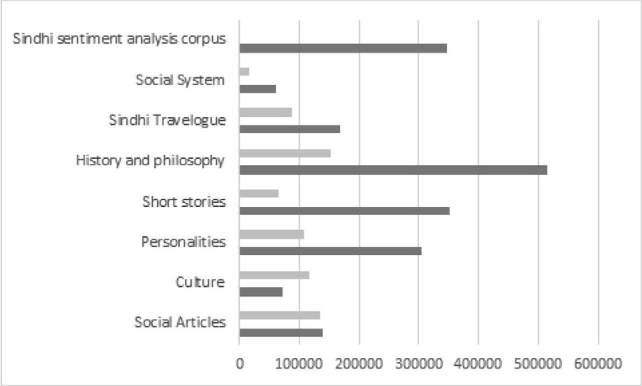


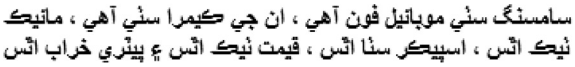
Fig. 2. Detail of Sindhi text corpora.

(Samsung mobile phone suthi phone aahay, in je camera suthi aahay, mic theek athas, speakers suthaa athas, qeemat theek athas aen in ji battery kharaab aahay). The meaning of the document of Sindhi sentiment based text corpus in English is “Samsung is a good mobile phone, its camera is good, mic is okay, speakers are good, price is okay and battery is bad.” The sentence presents the opinions and sentiments for mobile cell phone and its features. Sentence is consisted of positive, neutral and negative opinions, however; positive sentiments are in majority. Fig. 4 shows feature names, which are noun, opinions and sentiments, which are adjectives and the polarities of opinions and sentiments.

3.1.3. Machine readable corpus

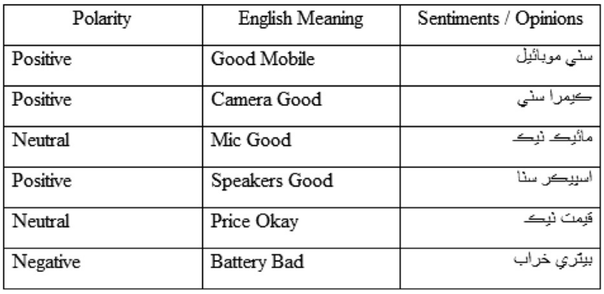
Advancement in computational technology enables people to understand the languages of the world. The role of computational linguistics and linguists are vital in this concern. Therefore, text corpus should be in machine readable form. Sindhi text corpus is available in machine readable form. The unicode utf-8 is utilized to recognized and read the Sindhi text corpus by machine. Plain Sindhi text corpus is processed on Sindhi NLP tools (http://www.sindhinlp.com) for tokenization, tagging, parsing, lemmatization, stemming and sentiment analysis and got better results. Fig. 5 shows the word tokenizatin, UPOS tagging and syntactic parsing of Sindhi text corpus document, Fig. 6 shows the sentiment analysis of Sindhi text. Sentiment analysis has been done on basis of polarity analysis. The results shows the sentiment analysis of text corpus document and presents the features of products along with opinion and sentiment of each feature separately. The polarity of whole document is shown on basis of high confidence level of overall polarity. Fig. 7 shows stemming and lemmatization process of text corpus document. Inflection makes Sindhi language complex for machines to understand. However, Sindhi NLP tools are trained to understand the inflected Sindhi text. The stemming and lemmatization processes are performed on Sindhi inflected text. Fig. 8 shows the stemming and lemmatization processes for inflected Sindhi text. Affixes and suffixes are removed from Sindhi inflected text by machine properly. Therefore, the discussing text corpus is very much suitable for different types of the analysis on Sindhi text.

These tools understand the Sindhi text and perform required computational linguistics and NLP operations on Sindhi text



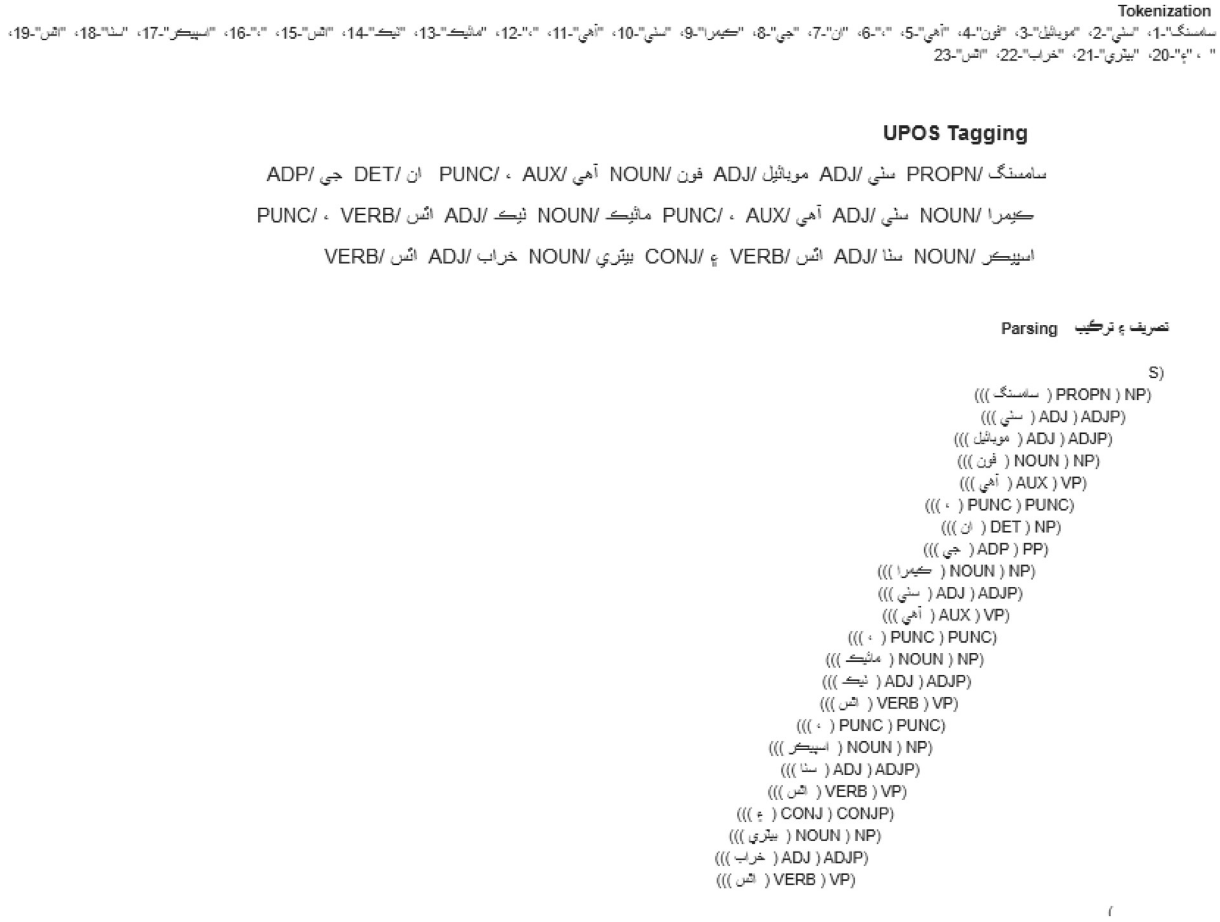
سامسنگ سٺي موبائيل فون آهي ، ان جي ڪيمرا سٺي آهي ، مائيڪ
ٺيڪ اٿس ، اسپيڪر سٺا اٿس ، قيمت ٺيڪ اٿس ۽ بيٽري خراب اٿس

Fig. 3. Sindhi Example text.



Polarity	English Meaning	Sentiments / Opinions
Positive	Good Mobile	سٺي موبائيل
Positive	Camera Good	ڪيمرا سٺي
Neutral	Mic Good	مائيڪ ٺيڪ
Positive	Speakers Good	اسپيڪر سٺا
Neutral	Price Okay	قيمت ٺيڪ
Negative	Battery Bad	بيٽري خراب

Fig. 4. Feature names, Sentiments and their polarities, identified from Sindhi text document.



Tokenization

"سامسنگ"-1، "سني"-2، "موبائيل"-3، "فون"-4، "آهي"-5، "،"-6، "ان"-7، "جي"-8، "ڪيمرا"-9، "سني"-10، "آهي"-11، "،"-12، "مائيڪ"-13، "ٺيڪ"-14، "اٿس"-15، "،"-16، "اسپيڪر"-17، "سٺا"-18، "اٿس"-19،
" ، "ءِ"-20، "بيٽري"-21، "خراب"-22، "اٿس"-23

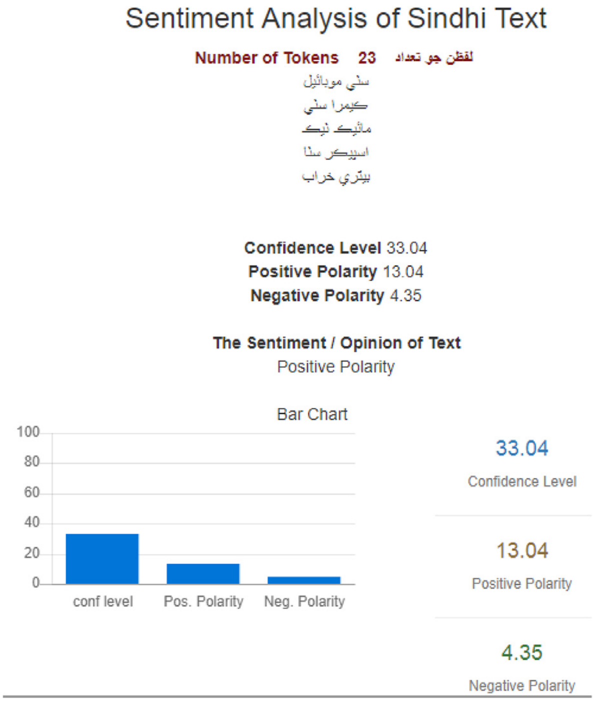
UPOS Tagging

سامسنگ /PROPN سني /ADJ موبائيل /ADJ فون /NOUN آهي /AUX ، /PUNC ان /DET جي /ADP
ڪيمرا /NOUN سني /ADJ آهي /AUX ، /PUNC مائيڪ /NOUN ٺيڪ /ADJ اٿس /VERB ، /PUNC
اسپيڪر /NOUN سٺا /ADJ اٿس /VERB ءِ /CONJ بيٽري /NOUN خراب /ADJ اٿس /VERB

تصريف ۽ ترڪيب Parsing

```
S)
((( سامسنگ ( PROPN ) NP)
((( سني ( ADJ ) ADJP)
((( موبائيل ( ADJ ) ADJP)
((( فون ( NOUN ) NP)
((( آهي ( AUX ) VP)
((( ، ) PUNC ) PUNC)
((( ان ) DET ) NP)
((( جي ) ADP ) PP)
((( ڪيمرا ) NOUN ) NP)
((( سني ) ADJ ) ADJP)
((( آهي ) AUX ) VP)
((( ، ) PUNC ) PUNC)
((( مائيڪ ) NOUN ) NP)
((( ٺيڪ ) ADJ ) ADJP)
((( اٿس ) VERB ) VP)
((( ، ) PUNC ) PUNC)
((( اسپيڪر ) NOUN ) NP)
((( سٺا ) ADJ ) ADJP)
((( اٿس ) VERB ) VP)
((( ءِ ) CONJ ) CONJP)
((( بيٽري ) NOUN ) NP)
((( خراب ) ADJ ) ADJP)
((( اٿس ) VERB ) VP)
(
```

Fig. 5. Word tokenization, UPOS tagging, syntactic parsing of Sindhi sentiment based text corpus document.



Sentiment Analysis of Sindhi Text

Number of Tokens 23 لفظن جو تعداد

سني موبائيل
ڪيمرا سني
مائيڪ ٺيڪ
اسپيڪر سٺا
بيٽري خراب

Confidence Level 33.04
Positive Polarity 13.04
Negative Polarity 4.35

The Sentiment / Opinion of Text
Positive Polarity

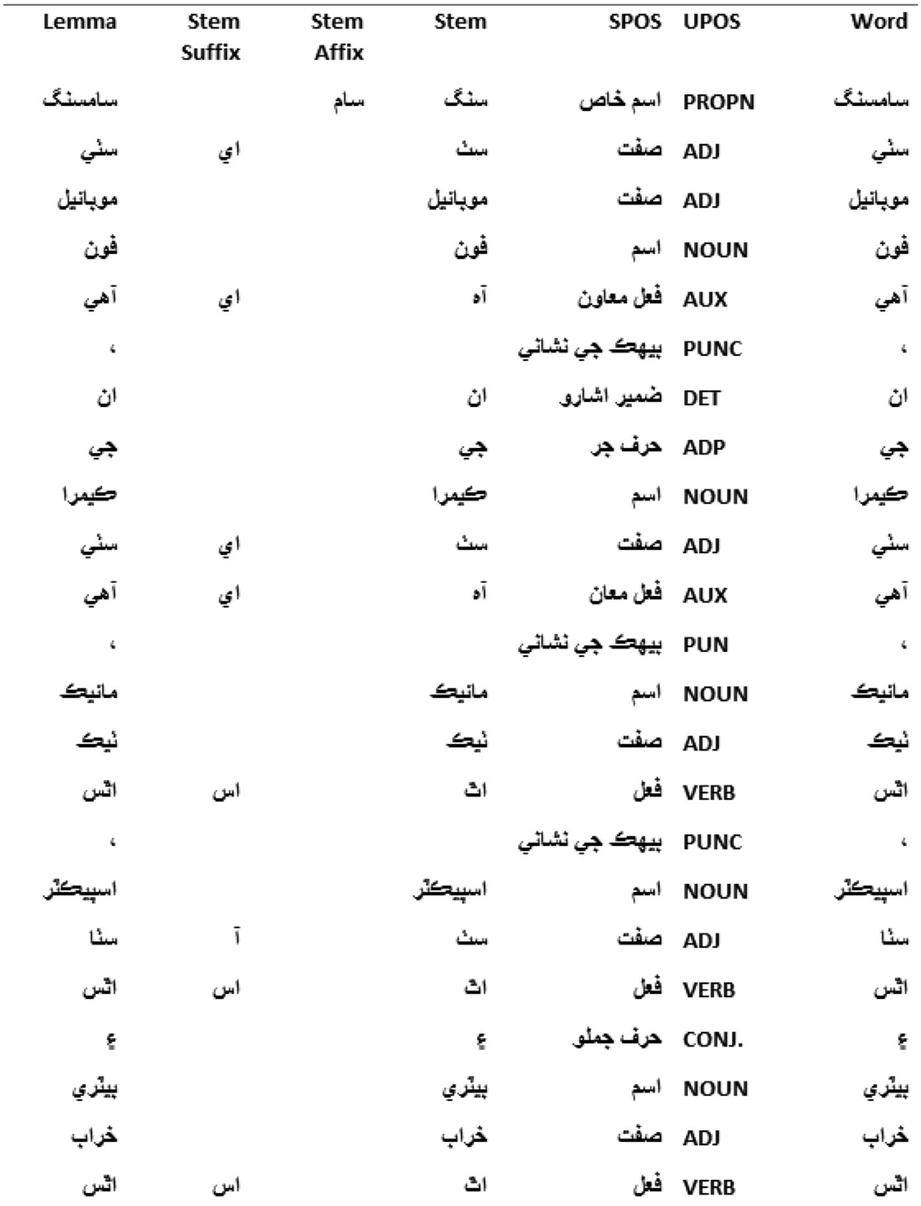
Fig. 6. Sentiment analysis process of Sindhi sentiment based text corpus document.

accordingly. The machine learning process may be performed on Sindhi text corpus for supervised and unsupervised analysis using different types of machine learning methods. The deep learning procedures may also be performed on Sindhi corpora for the deep analysis of Sindhi text. These all processes prove that Sindhi text corpus is in machine readable format completely.

3.2. Sindhi stop words

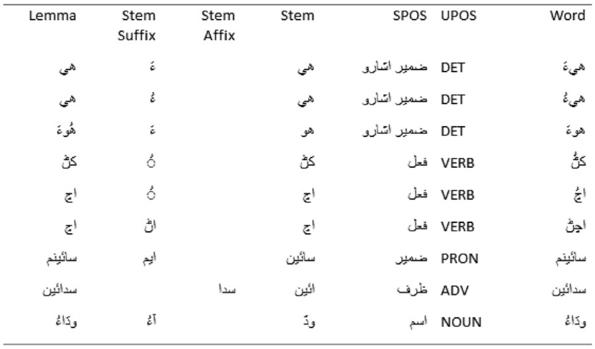
Stop words are not significant and important for information retrieval, search engines and other text analysis processes. The purpose of these words in sentence is to complete the sentence and to give a sense of understanding of sentence, so stop words construct the sentences properly. Computer science does not give importance to stop words but it filters these words during the searching and other text analysis processes. To process the Sindhi text corpus for analysis, Sindhi stop words are identified from described text corpus. Sindhi stop words are selected from prepositions or Ad-positions, determiners, verbs, conjunctions, interjections and articles. Fig. 9 describes some of the identified stop words from Sindhi text corpus to show the style and structure of Sindhi stop words.

Frequency of stop words in presenting sentiment based Sindhi text corpus is calculated properly to show the availability of Sindhi stop words in text corpus. Table 1 shows the stop words, their pronunciation, meaning in English, grammatical status and frequency



Lemma	Stem Suffix	Stem Affix	Stem	SPOS	UPOS	Word
سامسنگ		سام	سنگ	اسم خاص	PROPN	سامسنگ
سٺي	اي		سٺ	صفت	ADJ	سٺي
موبائيل			موبائيل	صفت	ADJ	موبائيل
فون			فون	اسم	NOUN	فون
آهي	اي		آه	فعل معاون	AUX	آهي
،				بيهڪ جي نشاني	PUNC	،
ان			ان	ضمير اشارو	DET	ان
جي			جي	حرف جر	ADP	جي
ڪيمرا			ڪيمرا	اسم	NOUN	ڪيمرا
سٺي	اي		سٺ	صفت	ADJ	سٺي
آهي	اي		آه	فعل معان	AUX	آهي
،				بيهڪ جي نشاني	PUN	،
مائيڪ			مائيڪ	اسم	NOUN	مائيڪ
نيڪ			نيڪ	صفت	ADJ	نيڪ
اٿس	اس		اٿ	فعل	VERB	اٿس
،				بيهڪ جي نشاني	PUNC	،
اسپيڪر			اسپيڪر	اسم	NOUN	اسپيڪر
سٺا	آ		سٺ	صفت	ADJ	سٺا
اٿس	اس		اٿ	فعل	VERB	اٿس
۽			۽	حرف جملو	CONJ.	۽
بيٽري			بيٽري	اسم	NOUN	بيٽري
خراب			خراب	صفت	ADJ	خراب
اٿس	اس		اٿ	فعل	VERB	اٿس

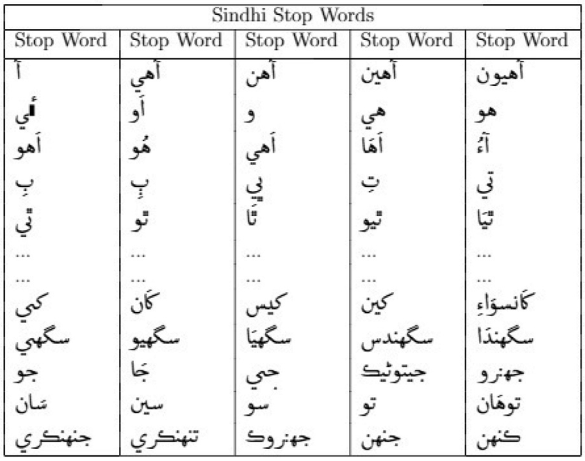
Fig. 7. Stemming and Lemmatization process for Sindhi sentiment based text corpus document.



Lemma	Stem Suffix	Stem Affix	Stem	SPOS	UPOS	Word
هي	ءَ		هي	ضمير اشارو	DET	هيءَ
هي	ءُ		هي	ضمير اشارو	DET	هيءُ
هُوءَ	ءَ		هو	ضمير اشارو	DET	هوءَ
ڪڻ	◌ُ		ڪڻ	فعل	VERB	ڪڻُ
اڄ	◌ُ		اڄ	فعل	VERB	اڄُ
اڄ	اڻ		اڄ	فعل	VERB	اڄڻ
سائينم	ايم		سائين	ضمير	PRON	سائينم
سدائين		سدا	ائين	ظرف	ADV	سدائين
وڏاءُ	آءُ		وڏ	اسم	NOUN	وڏاءُ

Fig. 8. Stemming and Lemmatization process for Sindhi inflected text.

of some stop words available in Sindhi sentiment based text corpus. The Sindhi word آهي (Aahay) means 'is', which is auxiliary verb (فعل معَاون), is used more than other stop words in Sindhi text



Sindhi Stop Words				
Stop Word	Stop Word	Stop Word	Stop Word	Stop Word
آ	آهي	آهن	آهين	آهيون
أي	اَو	و	هي	هو
اَهو	هُو	اَهي	اَهَا	آءُ
بِ	پِ	پِي	ٿِ	تِي
ٿي	تو	ٿَا	ٿيو	ٿيَا
...	...	...	...	...
...	...	...	...	...
ڪي	ڪَان	ڪيس	ڪين	ڪَانسوَاءِ
سگهي	سگهيو	سگهيَا	سگهندس	سگهندَا
جو	جَا	جي	جيتوڻيڪ	جهڙو
ٿان	سين	سو	تو	توهَان
جنهنڪري	تنهنڪري	جهڙوڪ	جنهن	ڪنهن

Fig. 9. Sindhi Stop Words.

Table 1
Sindhi Stop Words, their pronunciation, meaning and frequency.

Sindhi Stop Words	Pronunciation	UPOS Tag set	English Meaning	Frequency in Text Corpus
جي	Aahay	AUX	is	27236
اِن	Jay/Jee	ADP	of	14000
أي	In	DET	this/that/it	12920
: :	Ee	Particle/ADV	only	1320
::	: :	: :	: :	: :
اَٿس	Athas	VERB		11339
پر	Par	CONJ.	but	4797
ڀي	Bhee	Particle/ADV	may also	2549
بِ	Bi	Particle/ADV	also	1455
تَ	Ta	ADP		787

corpus. Therefore, Sindhi word آهي (Aahay) is ranked at high level in list of stop words, however, another Sindhi word جي (ji or jay) which means 'of' in English and it is ad-position, is comes at second highest level in Sindhi text corpus. Sindhi word اِن (In), which is determiner (ضمير اِشَارو) and means this or that in English, is ranked at third highest level. A Sindhi word اَٿس (Athas), which is verb (آهي) stands at fourth rank in Sindhi text corpus. The higher frequency of stop words shows the importance of verbs and prepositions or ad-positions in Sindhi text.

4. Results and discussions

The text corpus is preprocessed and normalized for the purpose of model development and performance analysis. All the identified stop words are removed from the text corpus. However, identification and removal of stop words process has been done carefully, because sometimes pronouns, especially personal pronouns (Campbell and Pennebaker, 2003) may be useful for text corpus analysis.

4.1. Document Term Matrix (DTM)

Now a day, text corpus analysis is important topic of natural languages process because several organizations focus on text corpus of different languages for multiple purposes. Sindhi language is written, read and spoken in several countries and Sindhi people write their views and posts on different products, personalities and topic on social media in Sindhi language as well as there are several blogs, which are written in Sindhi language, therefore, text corpus of Sindhi language is more important for different types of organizations, linguistics and NLP research. DTM presents the text vectors to show their usage, variation and feature distribution in different documents. It is like a grid of terms and documents, which shows the internal connection of both entities. This connection finds and fixes the terms in documents of text corpus, which is called feature distribution. The feature shows the occurrence of frequency of tokens in DTM. It is not necessary that all terms should be available in all documents of text corpus. The vector of all tokens frequencies is specified for document of text corpus, which is assumed as multivariate sample. Therefore, vectorization is term for converting the text corpus documents to numerical feature vectors.

The DTM for Sindhi text corpus is two dimensional matrix containing C columns and N rows. The columns show distinct words and rows show availability of those distinct words in documents. Each row in DTM presents the document of text corpus. The documents are shown with numbers like $0, 1, 2, 3, \ldots\ldots\ldots\ldots\ldots\ldots, n$. Each cell of DTM shows the availability and frequency of unique word in document. Generally, *xij* explains the value of the *jth* data variable for the *ith* document or row, where $i = 1, 2, 3, 4, \ldots\ldots\ldots\ldots\ldots\ldots, n$ and $j = 1, 2, \ldots\ldots\ldots\ldots\ldots\ldots, c$. Therefore, M = c x n, whose *ith* and *jth* elements are *mij*. Here, m_i is vector of length *c*, consisted of c data variable measurements for the *ith* document. The vectors of matrix are columns of matrix, whereas documents of matrix are rows of matrix.

N-gram model is used for information retrieval and several other functions of computational linguistics to make the language models, language feature analysis and etc. The grams are text items whereas n-grams find the adjoined items of n items from the given corpus. n-grams may be uni-gram, bi-gram and tri-gram. Uni-gram shows size of n-gram as one gram, bi-gram shows the size of n-gram as two grams and tri-gram shows size of n-gram as three grams. Therefore, n-grams of Sindhi text finds the sequence of Sindhi words available in Sindhi text corpus. For example, the uni-gram, bi-gram and tri-gram of a Sindhi sentence سنڌي ٻولي دُنيَا جي پرَاڻي ٻولي آهي (Sindhi language is oldest language of the world) are presented below. Sindhi text sentence is broken into uni-gram, bi-gram and tri-grams properly.

Uni-gram model
سنڌي
ٻولي
دُنيَا
جي
پرَاڻي
ٻولي
آهي
Bi-gram model
سنڌي ٻولي
ٻولي دُنيَا
دُنيَا جي
جي پرَاڻي
پرَاڻي ٻولي
ٻولي آهي
Tri-gram model
سنڌي ٻولي دُنيَا
ٻولي دُنيَا جي
دُنيَا جي پرَاڻي
جي پرَاڻي ٻولي
پرَاڻي ٻولي آهي

Therefore, DTM development is used to know the frequency and variation of Sindhi terms in separate documents of text corpus.

This shows the features and importance of Sindhi language and lexicons. DTM for Sindhi text corpus is developed using n-gram model where n = 2, therefore, frequency of words is associated with documents available in text corpus on basis of n-gram words. Extraction of 2gram shows the complexity of Sindhi language. It is important feature of Sindhi language text corpus that it uses compound words in several documents of the corpus. The frequency of uni-gram words may be common but frequency of 2-gram words is not common. The correspondence of the 2-gram terms to documents shows the importance of Sindhi text corpus for text mining and analysis. DTM sows the language variation because dissimilar terms of Sindhi text corpus are identified by it. Fig. 10 shows the Document Term Matrix of Sindhi sentiment based text corpus. The matrix shows the availability of terms in separate document and the language variation. There are separate sentiments used for separate terms which confirm the language variation of Sindhi language. DTM shows the frequency of Sindhi compound words, which are observed on basis of 2-gram model. The words are consisted of nouns and adjectives. Adjectives shows the polarity of Sindhi noun words.

The DTM shows results with large number of terms and their availability in different documents of Sindhi text corpus. It classifies and recognizes the quantity of motivating and interesting features of documents of text corpus. DTM presents the frequency and co-relation of terms in separate documents of text corpus, which shows the significance of document terms. These terms may be vital for information retrieval, sentiment analysis, cluster analysis and other supervised and unsupervised machine learning processes. The matrix presents the language variation because terms vary in different documents of text corpus.

4.2. Term frequency inverse document frequency (TF-IDF)

The term weighting schemes are novel and significant for information retrieving process, therefore, the main function of term weighing is to measure the salience feature of term of document (Paik, 2013) therefore, documents are ranked on basis of weight of terms. Term frequency inverse document frequency, also pronounced as tf-idf is significant and useful statistical model for information retrieval, text mining and machine learning processes. It finds-out the key terms from all documents of text corpus to make them vital to a document. The tf-idf weight describes the importance of a word, utilized as term, to a document in text corpus. Basically, tf-idf is combination of two terms: TF (Term Frequency) and IDF (Inverse Document Frequency), thus, it is two-dimensional matrix. Term Frequency (TF) counts the frequency of term or terms or word or words available in a documents of text corpus, therefore, the number of frequency of term divided by total number of terms in document presents the total term frequency of word in document. Following equation shows the method of term frequency measurement.

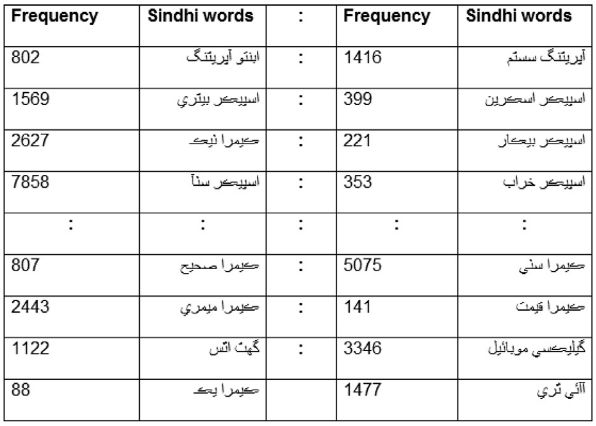
$$f_{t,d}/\sum_{t\prime\in d} f_{t\prime,d}\sum_{t\prime\in d} f_{t\prime,d} \tag{1}$$

The frequency of term t in document d shows the total number of frequencies of term/terms in document d. TF equation describes the number of frequencies of term t, appears in document d divided by total number or sum of terms available in document d. The IDF shows the status of word that either it is distinct or common, therefore, it presents the word with all its information. IDF is measured as logarithm of the total number of documents available in text corpus.

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

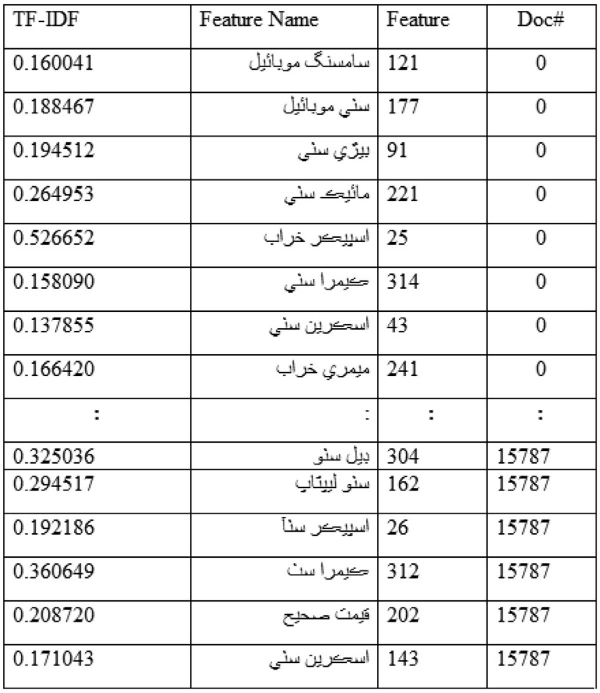
The IDF (t) is measured as the logarithm of number of documents d available in the text corpus N divided by number of frequencies of documents with term t available in all documents. TF-IDF uses N-gram model. The numerator N shows the number of documents available in corpus N, whereas, denominator describes the total number of frequencies of documents in which terms t are available. The document d belongs to text corpus therefore, D shows the number of documents d (d1, d2, d3,..., dn) in text corpus of Sindhi. TF-IDF of Sindhi text corpus finds out the Sindhi significant words which perform important role in documents of text corpus. The feature names, which are especial terms, are different from each other and significant for documents of text corpus. The feature names are derived from terms available in Sindhi text corpus. The feature shows the N-gram terms. Fig. 11 shows the TF-IDF of document terms available in the Sindhi text corpus.

Results of TF-IDF show the importance of terms and documents available in Sindhi text corpus. The tf-idf shows the value of word to a document of text corpus, therefore, tf-idf of Sindhi text corpus



Frequency	Sindhi words	:	Frequency	Sindhi words
802	ابنٽو آپريٽنگ	:	1416	آپريٽنگ سسٽم
1569	اسپيڪر بيٽري	:	399	اسپيڪر اسڪرين
2627	ڪيمرا ٺيڪ	:	221	اسپيڪر بيڪار
7858	اسپيڪر سٺآ	:	353	اسپيڪر خراب
:	:	:	:	:
807	ڪيمرا صحيح	:	5075	ڪيمرا سٺي
2443	ڪيمرا ميمري	:	141	ڪيمرا قيمت
1122	گهٽ اٿس	:	3346	گليڪسي موبائيل
88	ڪيمرا يڪ		1477	آئي ٽري

Fig. 10. DTM presents Sindhi terms and their frequency in text corpus documents.



TF-IDF	Feature Name	Feature	Doc#
0.160041	سامسنگ موبائيل	121	0
0.188467	سني موبائيل	177	0
0.194512	بيڙي سني	91	0
0.264953	مائيڪ سني	221	0
0.526652	اسپيڪر خراب	25	0
0.158090	ڪيمرا سني	314	0
0.137855	اسڪرين سني	43	0
0.166420	ميمري خراب	241	0
:	:	:	:
0.325036	بيل سنو	304	15787
0.294517	سنو ليپٽاپ	162	15787
0.192186	اسپيڪر سنآ	26	15787
0.360649	ڪيمرا سٺ	312	15787
0.208720	قيمت صحيح	202	15787
0.171043	اسڪرين سني	143	15787

Fig. 11. TF-IDF of Sindhi terms and their association with documents.

shows the results of words which are derived through 2-gram model from Sindhi text corpus documents. The number of frequencies of 2-gram words to documents of text corpus show the feature of Sindhi terms. The presented features and their names show the correspondence to different documents of the Sindhi text corpus. These features of documents are very much significant and vital to the documents of text corpus.

Text analysis is important topic of data mining applications and research because internet resources produce large amount of text of social, political, educational, scientific text and etc. Analysis of text corpus enable organizations to extract useful data and information, therefore, decision makers may take good decisions, translators may translate the language to other languages as well as language variation and feature distribution may be observed for the purpose of information retrieval.

5. Conclusion

Sindhi text corpus is provided with additional linguistic features. The analysis is performed on text corpus and its features. The evaluation and analysis is performed on a single plain text corpus. The enough number of complex words make the Sindhi text corpus rich. The corpus is prepared for Document Term Matrix development using N-gram model. DTM classifies and recognizes the Sindhi text terms and shows frequency of them in different documents. DTM presents the language variation in form of language terms and topics. Term frequency inverse document frequency matrix shows better results using N-gram model. TF-IDF shows the importance of word to document of text corpus. DTM and TF-IDF show the significance of Sindhi text corpus for information retrieval, sentiment analysis, and pattern recognition.

Research studies on different topics have brought change in computer science, applied science, social science and other domains, therefore, this is continuous process of making the things perfect as well as beneficial for the development of society. This research study is the basic research study on development and analysis of Sindhi text corpus using Arabic-Persia script, however, more research work is required for the analysis of Sindhi text corpus using Word2vec, cluster analysis, term similarity analysis, topic modeling and sentiment analysis. Study contributes Sindhi text corpora to NLP and computational linguistics fields for future research.

Acknowledgment

The research study is produced from my research work on "Sentiment Analysis for Sindhi Text" which is submitted at SZABIST Karachi Sindh Pakistan as my doctoral research. I acknowledge the support of Program Coordinator Dr. Hussnain Mansoor Ali Khan and head of department Dr. Imran Amin for provision of lab resources and environment.

References

Alana, G.A., 2010. Sindhi Boli jo Tashreehi grammar. Sindhi Language Authority, Hyderabad, Sindh Pakistan.

Dootio, M.A., Wagan, A.I., 2019. Syntactic parsing and supervised analysis of sindhi text. Elsevier J. King Saud Univ.-Comput. Inf. Sci. 31, 105–112. https://doi.org/10.1016/j.jksuci.2017.10.004.

Ali, M., Wagan, A.I., 2017. Sentiment summerization and analysis of sindhi text. Int. J. Adv. Comput. Sci. Appl. 8 (10), 296–300.

Jumani, A.K., Memon, M.A., Khoso, F.H., Sanjrani, A.A., Soomro, S., 2018. Named entity recognition system for sindhi language. In: International Conference for Emerging Technologies in Computing. Springer, pp. 237–246.

Shah, S.M.A., Ismaili, I.A., Bhatti, Z., Waqas, A., 2018. Designing xml tag based sindhi language corpus. In: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) IEEE, pp. 1–5.

Ali, M., Wagan, A.I., 2019. An analysis of sindhi annotated corpus using supervised machine learning methods. Mehran Univ. Res. J. Eng. Technol. 38 (1), 185–196.

Dootio, M.A., Wagan, A.I., 2018. Unicode-8 based linguistics data set of annotated sindhi text. Elsevier Data in Brief 19, 1504–1514.

Kennedy, G., 2014. An Introduction to Corpus Linguistics. Routledge.

Bosco, C., Patti, V., Bolioli, A., 2013. Developing corpora for sentiment analysis: the case of irony and senti-tut. IEEE Intell. Syst. 28 (2), 55–63.

Schäfer, R., Bildhauer, F., 2013. Web corpus construction. Synthesis Lectures on Human Language Technologies 6 (4), 1–145.

Das, A., Bandyaopadhyay, S., Gambäck, B., 2012. The 5w structure for sentiment summarization-visualization-tracking. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 540–555.

Agrawal, S.S., Abhimanue, S.B., Bansal, S., Mahajan, M., 2014. Statistical analysis of multilingual text corpus and development of language models. In: LREC. Citeseer, pp. 2436–2440.

Baseer, F., Habib, A., Ashraf, J., 2016. Romanized urdu corpus development (rucd) model: Edit-distance based most frequent unique unigram extraction approach using real-time interactive dataset. In: Innovative Computing Technology (INTECH), 2016 Sixth International Conference on, IEEE, pp. 513–518.

Motlani, R., 2016. Developing language technology tools and resources for a resource-poor language: Sindhi. Proceedings of NAACL-HLT, 51–58.

Rahman, M.U., 2010. Towards sindhi corpus construction. In: Conference on Language and Technology, Lahore, Pakistan.

Krishna, B.V., Pandey, A.K., Kumar, A.S., 2018. Feature based opinion mining and sentiment analysis using fuzzy logic. In: Cognitive Science and Artificial Intelligence. Springer, pp. 79–89.

Negi, S., Buitelaar, P., 2017. Suggestion mining from opinionated text. Sentiment Anal. Soc. Networks, 129–139.

Campbell, R.S., Pennebaker, J.W., 2003. The secret life of pronouns: flexibility in writing style and physical health. Psychol. Sci. 14 (1), 60–65.

Paik, J.H., 2013. A novel TF-IDF weighting scheme for effective ranking. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval ACM, pp. 343–352.