REPERTUS
JOURNAL OF LINGUISTICS,
LANGUAGE PLANNING & POLICY
ISSN 2791-1934

Researching on Analysis and creating Corpus from Primary level Sindhi language Book for Sindhi

Naveen Talpur [a]
Mir Jahanzeb Talpur [b]
Timotheous Samar [c]

[a] Mehran university of engineering and Technology, Sindh Pakistan
[b] Mehran university of engineering and Technology, Sindh Pakistan
[c] Mehran university of engineering and Technology, Sindh Pakistan

KEYWORDS
Sindhi Corpus, UOPS, Sentimental analysis, Document Term Metrix

ABSTRACT

Sindhi is an amusing vernacular with a large abundance of pieces of literature and non-literary works. Despite the availability of several books, newspapers, magazines, and internet resources for developing Sindhi text corpora, a suitable and effective textual corpus could not be generated and offered accessible for investigation, language characteristics research, semantics assessment, and information gathering systems. The paucity of tools for computational linguistics research and NLP apps for Sindhi is stimulating complications at this time. Moreover, we have built Sindhi text libraries to provide computer linguistics, NLP specialists, and academics with text resources. The Sindh Text Book Board and primary school textbooks are used to create the Sindhi text corpus. Using the 2-gram approach of the n-gram model, using the Document Term Matrix and TF-IDF models, a Sindhi belief text dataset is produced and evaluated. The dataset might be useful for research on linguistic suggested work, topic detection, and sentiment classification by aspect.

1. Introduction

Language is made up of a collection of symbols that may be utilized in both written and spoken forms of communication. It is a fundamental and important resource for human society's communication and commercial transactions. Languages allow people to exchange their ideas, values, and resources. Language, like other human difficulties, is an evolutionary process and a problem (Alana, 2010). In our digital age, social media plays an important role. The digital society is propelled by information technology infrastructures, social media technologies, and artificial intelligence. People interact with one another via social media networks, hence social media are vast networks of relationships. On social media networks, users convey their emotions, feelings, sentiments, and thoughts about shared entities (Alejandro, 2013, Christopher, 2018). People are giving their opinions in English and their native languages, thus NLP researchers and computational linguists have a research platform to study many elements of language and product evaluations. The connection between communication and culture may be seen in our daily interactions between people and groups. Our language must have been impacted by our location as well as the ethnic areas around us.

In this example, culture refers to the human lifestyle. People learn, thought, experience, think, and pursue what is culturally suitable. Language, relationship, tradition, communications practice, social action, economic activity, politics, and technology were all founded on the cultural pattern. Javanese, Malay, and English are the languages spoken. This is because they were born or nurtured in a culture that included these components. What people do, and how they behave, is a reaction to cultural functions (Porter & Samovar, 2006).

Only a few documents have been utilized and are available in Roman script on mobile phone devices, mobile phones, and internet communications. Unfortunately, the linguistic corpus and extensive computational lexicon have yet to be launched, despite the fact that they are critical for the creation of Sindhi language processing resources. It is a fact that a large amount of written material in Sindhi is available both offline and online. Sindhi Corpus' script is Persio-Arabic, and it was created in Persio-Arabic using UTF-16 coding.

Sindhi is one of Pakistan's main languages, spoken by 30-40 million people (Collie, 2006). Sindhi is a popular language on the internet. Sindhi websites, literary sites, news websites, and comment sections are growing in popularity. Sindhi is the second most frequently spoken engraved language in Pakistan, behind Urdu. Regardless of its widespread use and popularity online, NLP researchers have access to just a few language processing tools, such as lexicons, typefaces, and rudimentary word processors. Tools for processing Sindhi such as linguistic corpora and a complete computational lexicon have yet to be developed. Persio-Arabic, Devnagri (Û), and roman (Sindhi) scripts are used to write Sindhi. Sindhi writings in Pakistan and India are written in the Persio-Arabic script. In India, Sindhi writing is also done in Devnagri script. The Roman alphabet is also gaining popularity (though it is not yet standardized). Only a few written papers are accessible in roman manuscripts, but it is widely used for online, mobile phone, as well as other wearable technology interactions. Because the majority of Sindhi written content is available in Persio-Arabic script both online and offline, The Sindhi corpus is being built in Persio-Arabic script.

The majority of the community of this state uses Sindhi not only for texting in mobile function today but also for text communications in applications, letters, etc. (Alejandro, 2013, Christopher, 2018) Cell phones, televisions and laptops acquire novice elements of our lives. Communication via computers and mobile phones, such as short messaging services (SMS), and applications such as Twitter, What's App, and Facebook, have increased significantly. Although

English is frequently used in the above-mentioned types of services, in many countries people adopt to correspond in their native language rather than English of course, the native language is a dynamic fount of contact. Therefore, researchers in Pakistan and India are focusing on the problem of local languages (Urdu, Sindhi, Punjabi, Hindi, etc.).

Information Retrieval 'IR' Data mining (DM) is important for making associations among words in a sentence or survey. Subject to analysis, outcome analysis and sentiment analysis are included in the NLP knowledge base. NLP characteristics are described in various ways, including B. Stems, lemmas, tokens, part-of-speech tagging (POS), stop words, shallow parsing, and named entity recognition (NER) cover important values in all NLP systems (Riaz, 2010). Although much effort has abided in English, Urdu and Arabic, there are still large gaps in the research and study of Sindhi. Building applications in your native language is essential for language survival and effective communication in your native language. Therefore, this article will focus on Sindhi.

These advancements and research projects serve as impulses for technologically underdeveloped languages across the world, such as Sindhi. Currently, a study is being conducted on the Sindhi language to assess and examine its linguistic issues (Dooti & Wagan et al., 2019). The aim of this study is to present the impetus to work on the sentimental analysis of primary-level Sindhi text from primary-level Sindhi subject books and construct a machine-readable corpus on Primary level Sindhi text.

2. Literature Review

Multilingual countries make it easier to understand and speak several languages, but computers struggle to comprehend the diversity of natural languages. Unicode overcomes linguistic challenges by making them understandable to computer systems. Proper corpus development necessitates the use of appropriate tools and methodologies. Bosco et al. (2013) explain that the production of a text corpus is divided into three stages: collecting, explanation, and analysis. The annotation corpora are appropriate for sentiment categorization. Although practical exams of language may be conducted using that language's corpus, the framework of any language can be a momentous and crucial aspect of a language on which studies are carried out or continuing. A corpus is a large collection of written text created intended for the determination of linguistic scrutiny (Kennedy, 2014). As a result, it is crucial information that offers dictionaries, linguists, and other specialists have a solid mastery of a language Corpus analysis offers data on morphology, syntax parsing, lexicon development, interpretation, dialectic, and other linguistic topics. Agrawal et al., (2014) conducted a study to understand the intricacies and variations across languages and use a multilingual corpus. They explain these findings by claiming that Nepali is more difficult to learn than Hindi or Punjabi The creation and analysis of Urdu script corpora, and as a result, they employed the K-means machine-learning technique to train the machine for cluster analysis, with favorable results (Baseer et al. 2016)

Dooti & Wagan (2018) focused on NLP and stated in their research that the Sindhi script has a wide range of classes and features. There is a lot of work done in English script, and there are NLP tools accessible in English scripts that can complete all jobs in English script; however, there is no robust application for feature extraction and corpus available in Sindhi. Similar to Arabic and Urdu, Sindhi is a right-handed writing system (Awan et al, 2018). However, the use of Sindhi script is expanding across all platforms, particularly social media. Textual interaction is being employed in a variety of venues, including online publications, newspapers, poetry, and websites for learning Sindhi. It indicates that a significant volume of data is accessible at many website domains. At this time, there are no online NLP tools that can execute tasks like tokenizing manuscripts into sections, phrases, words, and letters. The authors of this study identified the elements of speech used in Sindhi text, but the most crucial work was sentiment analysis of Sindhi text across several contexts.

The Roman script is being used worldwide, much as English is increasingly used in information technology and computer science fields. The Roman script is simpler to write in than other scripts like Sindhi, Urdu, Arabic, etc., hence the bulk of individuals prefer it to online platforms to communicate. For this script, new applications have been created by researchers. Comparing Roman script with Arabic script presented writers with an additional challenging scenario while writing in Arabic. In terms of writing scripts, the Arabic language has characteristics of Sindhi (Hakro,2014)

These studies demonstrate the significance of corpora in language modelling and learning. Sindhi is a morphologically and grammatically opulent and complicated language, therefore creating a manuscript body and examining its contents is essentially solving computational linguistics issues in Sindhi. At certain levels, the organization of appropriate native texts is

Comparable, but there are distinctions in sounds, vocabulary formations, structural frameworks, and sentence construction; hence, research and some linguistic approaches may be effective at some levels but not totally. Sindhi is a structurally rich and complicated language since it uses all sorts of morphology in its text, including phrase terms and complicated verbs As a consequence, this study employed a 2-gram model to develop DTM and TF-IDF models for detecting difficult and complicated phrases in the Sindhi text corpus (Alana, 2010).

Sindhi language processing resources, aside from typefaces, console strategy, and a few electronic glossaries really aren't open to the public (CRULP, 2010). There are no studies or development efforts underway for resources such as language corpora and a full computer vocabulary. Multiple research individuals and organizations are striving to generate linguistic corpora for various Pakistani languages.

The Hindi and Punjabi corresponding body created by CDAC Noida is another intriguing linguistic dataset. Corresponding datasets do not exist for Sindhi, Balouchi, Siraiki, or numerous other Pakistani languages. Contrary to certain other Pakistani languages (with the exception of Urdu), Sindhi material in digital form is commonly accessible and is constantly being gathered for the corpus under consideration.

3. Research Methodology and Theoretical Framework

Text corpus development is carried out utilizing approaches from the manuscript body construction technique. The passages were obtained from the Primary Sindhi subject books published by the Sindh Text Book Board. In order to provide a consistent corpus for NLP researchers, this work created a new dataset for the Sindhi language with the help of the Quantitative method. Using UPOS, morphological tokens, emotive analysis, and Document Term Metrix, an annotation procedure is carried out. Explain the syntax of sentences of the Sindhi text corpus, including procedural approaches and data gathering sources. The procedure begins with a thorough grasp of the problem and concludes with the creation of a text corpus.

3.1 Sindhi language corpus development:

Character support and a Character encodings Sindhi keyboard design (bhurgari, 2010) have increased the accessibility of Sindhi writing in Unicode on the computer. Key Sindhi corpus is motivated by a number of factors. Development is the availability of Sindhi writing on the internet Newspapers, blogs, literary websites, and forums are all good places to start. Regardless of the fact that it is available on the internet, Materials don't have a lot of text, but they do have a lot of it growing by the day, and a corpus is being gathered constantly. Preprocessing software procedures Normalization, tokenization, and frequency computation are all steps in the process that are written in C# and run on the Microsoft.net foundation libraries.

3.2 Corpus acquisition:

Data is taken from Sindh Text Book Board elementary school textbooks in Sindhi, which cover science, social studies, Islamiyat, Sindhi, and mathematics. Pakistan research, talks, and viewpoints are among the several subdomains.

3.3 Normalization:

Although almost all of the data was already in Unicode format, all of the gathered material was transformed to normal UTF-16 encoding. Several Unicode points are used to represent letters, and It is typical for identical representations of composite and deconstructed forms to be reduced to the same procedure (Sarmad, 2008). When dealing with text processing, letters with aspirated variants, such as, which are combinations of two Unicode characters, are regarded as single letters.

3.4 Machine-readable corpus:

People can now comprehend the languages of the globe thanks to advances in computing technology. In this regard, computational linguistics and linguists play a critical role. As a result, the textual corpora ought to be computer accessible. A machine-readable Sindhi language corpus is accessible. The Sindhi text corpus is identified and read by machine using the Unicode utf-8. Encoding, labelling, parser, tokenizing, stemmed, and sentiment analysis was performed on a plain Sindhi text corpus using Sindhi NLP tools (http://www.sindhinlp.com) with superior results.

3.4.1 UPOS Tagging

NOUN/ .اسين / آهي PRON/ ڳوٺ NOUN/ کي ADP/ صاف ADJ/ رکندا / سٽرو VERB/ .آهيون / / /
NOUN/ ڏاڍو ADJ/ سٺو ADJ/ کي / آهي.اسان ADP/ ڳوٺ / پنهنجي NOUN/ سان ADP/ ڏاڍو ADJ/ پيار
اسان PRON/ جو ADP/ ڳوٺ

3.4.2 Tokenization

اسان"-1، "جو"-2، "ڳوٺ"-3، "ڏاڍو"-4، "سٺو"-5، "آهي.اسان"-6، "کي"-7، "پنهنجي"-8، "ڳوٺ"-9، " اسان
ڏاڍو"-11، "پيار"-12، "آهي."-13، "اسين"-14، "ڳوٺ"-15، "کي"-16، "صاف"-17، "سٽرو"-"، 10-"
" ،رکندا"-19، "آهيون."-20، ""-21، ""-22"، 18

3.4.3 Parsing تصريف ۽ ترڪيب

```
S)
        ((( اسان ) PRON ) NP)
                        ((( جو ) ADP ) PP)
                        ((( ڳوٺ ) NOUN ) NP)
                        ((( ڏاڍو ) ADJ ) ADJP)
                        ((( سٺو ) ADJ ) ADJP)
                        ((( کي ) ADP ) PP)
                        ((( ڳوٺ ) NOUN ) NP)
                        ((( سان ) ADP ) PP)
                        ((( ڏاڍو ) ADJ ) ADJP)
                        ((( پيار ) NOUN ) NP)
                        ((( اسين ) PRON ) NP)
                        ((( ڳوٺ ) NOUN ) NP)
                        ((( کي ) ADP ) PP)
                        ((( صاف ) ADJ ) ADJP)
                        ((( رکندا ) VERB ) VP)
```

Statistical Analysis of UPOS tagging and Syntactic Parsing
Number of Tokens 20 لفظن جو تعداد
Execution Time: 0.03843 s وقت

Table:3.1: Phrase tagging to Sindhi text

Phrase	Total Number of Tagged Phrases	Percentage
Noun Phrase	6	30.00
Adjective Phrase	4	20.00
Adverbial Phrase	0	0.00
Verb Phrase	1	5.00
Proposition Phrase	4	20.00
Interjection Phrase	0	0.00

Table:3.2: UPOS tagging to Sindhi text

UPOS	Total Number Tagged	Percentage
Nouns	4	20.00
Proper Nouns	0	0.00
Pronouns	2	10.00
Determiners	0	0.00
Verbs	1	5.00
Auxiliary verbs	0	0.00
Adjectives	4	20.00
Number Adjectives	0	0.00
Adverbs	0	0.00
Appositions	4	20.00
Conjunctions	0	0.00
Participles	0	0.00
Interjections	0	0.00
Unknowns	0	0.00
Punctuation	0	0.00
Symbols	0	0.00
Periods	0	0.00

Table:3.3 Analysis of Sindhi Morphological Tokens

Morphological Words	Total Number in text	Percentage
Simple Words	13	65.00
Complex Words	2	10.00
Compound Words	0	0.00
Reduplicated Words	0	0.00

These tools comprehend Sindhi text and execute the necessary computational linguistics and natural language processing operations on it. Machine learning may be used for the Sindhi source text for unsupervised and supervised analysis utilizing various machine learning algorithms. For the deep study of Sindhi text, deep learning processes may be applied to the Sindhi corpus. All of these procedures demonstrate the Sindhi text corpus is totally machined and accessible.

4. Sentimental analysis:

Words are indeed the categorization of human experimentation and encounters. What we notice, attend to, sense, witness, and behave in various ways are all influenced by our beliefs, conceptions, and experiences. When someone speaks or writes, he or she uses the same terms depending on his or her point of view and judgment. In light of all of this, the following are some words to consider. Although the same word is provided in many senses so that the meaning of the term may be readily understood and judged throughout the narrative. In Sindhi, consonants and inflectional can be removed from the concatenated text by using a machine the right way (Rahman, 2009).

Number of Tokens 20 لفظن جو تعداد
Confidence Level 45
Positive Polarity 25.00
Negative Polarity 0.00

The Sentiment / Opinion of Text
Positive Polarity
Bar Chart
45 Confidence Level

25.00
Positive Polarity

0.00
Negative Polarity

Text analysis is an essential issue in data mining applications and research because intellectual text and academic, governmental, and social text analysis are online resources that generate vast amounts of text. Organizations collect relevant data and information to evaluate the text corpus, making it simpler to interpret the languages and for policymakers to make appropriate judgments.

4.1 Document term matrix:

Text corpus research is a hot issue in natural language processing these days, with numerous groups focusing on text corpora of various languages for a variety of reasons. The Sindhi language is authored, perused, spoken and in many countries, and Sindhi people are posting their opinions on various goods, people, and issues on social media in Sindhi, as well as various other languages blog sites authored in the Sindhi language. As a result, the Sindhi language text corpus is more important for different sorts of companies, linguistics, and NLP research. The processing elements are shown in DTM to demonstrate their use, variety, and feature distribution in various publications. It's essentially a grid of terminology and documents that indicate how the two companies are connected inside. Feature distribution is the process of finding and fixing words in documents from a text corpus. The incoming visual is the frequency with which tokens appear in DTM.

The N-gram model has been utilized in data acquisition as well as a range of other computer linguistics applications such as language modelling and pattern evaluation. Although the grams are language pieces, the n-grams find the nearby items of n items in a corpus. N-grams are classified into three types: unigrams, bigrams, and trigrams. The uni-gram represents n-gram as one gram, the bi-gram as two grams, and the tri-gram as three grams. As a result, Sindhi text n-grams discover the series of Sindhi terms in the Sindhi-given text. For Example: -1-"اسان"، 2-"جي"، 3-"ڳوٺ"، 4-"۾"، 5-"ڇوڪرن،ڇوڪرين"، 6-"لاء"، 7-"اسڪول"، 8-"آهي".

Uni-Gram

اسان

جي

ڳوٺ

"۾"-

لاء

Bi-Gram

چوڪرين،چوڪرن

آهي اسڪول

Tri-Gram

ڳوٺ جي اسان

لاء چوڪرين،چوڪرن

As a result, the construction of DTMs is used to determine the prevalence and difference of Sindhi words in various text corpus sources. This exemplifies the structures and relevance of Sindhi lexemes and languages. The n-gram model was used to build the DTM for the Sindhi document collection, with n = 2. As a consequence, using n-gram words, the frequency of words is connected with the texts in the provided text. The extraction of 2 grams demonstrates the complexities of the Sindhi language. The employment of compound terms in various texts in the Sindhi language text corpus is an essential characteristic.

5. Conclusion

Many topics have changed significantly as a result of research inquiries in the fields of social science, industrial science, computer programming, and other areas. It's a never-ending process of refining things for the sake of societal advancement. The fundamental research study focuses on the analysis and development of Sindhi textual data. The Arabic-Persia script is employed for this purpose, and additional study is obligatory and intended for the examination of the Sindhi text corpus. Word 2 Vic, sentiment analysis, topic modelling, and cluster analysis are all utilized for this aim. In Sindhi text corpora, computational linguistics and natural language processing (NLP) are being contributed for future research.

The Sindhi corpus development effort is critical for language processing in the absence of Sindhi language sources. Despite the corpus's size and first output, its current position will serve as a foundation for future Sindhi language research because it is a natural language process. The script frequency, which also includes repetitive motion and word embedding, provides a foundation for compact keyboard design and effective text processing in smart gadgets and cell phones. Unigram and bigram frequencies at the word level provide foundation apps for spelling clarification and robotic sentences. Further corpus augmentation will be especially beneficial for higher-level receptive language goals including knowledge discovery and extracting, translation software, text analytics, grammatical assessment, and properties of language.

This paper presents the creation and examination of the corpus of the Sindhi dialect essential level course reading and the complexities and challenges of analyzing the Sindhi dialect. This work too reveals the significance of the Sindhi dialect within the NLP world. The corpus of content has been created and discharged. The as it were impediment of this corpus is that content is taken from an essential level lesson one book. As each period has its claim content, it can be encouraged and amplified by counting information from distinctive periods. In expansion to that the corpora creation pipeline can be computerized in future work

5.1 Recommendation for Future Research:

The corpus is continuously actuality possessed and as a result being corrected. Currently, the corpus is commonly based on machine learning encrypted primary-level Sindhi text. Studies are underway to appropriate annotations, POS markup, text-based vocabulary advancement and classification of documents based on n-grams. The Sindhi token algorithm needs to be implemented. Missing basic end-of-sentence punctuation marked in Sindhi; semicolon and alternative interval used to end a sentence in Sindhi texts. The sentence segment is another important area to work on. More definite Computer Sindhi language study is required to further development and maturation of the text block. Because the current example has no complete point of sale Tagging algorithms are available for Sindhi. The current POS tagging data applicable to Sindhi needs to analyze and dig deeper. Need a set of Sindhi beacons intended before the PLV markup of the text. Improve qualitative, quantitative, and relevant Full annotation and statistical analysis are areas that need to be developed in depth.

References

Allana, G. A. (2010). *Sindhi Boli jo tashreehi grammar* (A detailed grammar of the Sindhi language)(Vol. 1). *Sindhi Language Authority (SLA), Sindh, 71000*.

Dootio, M. A., & Wagan, A. I. (2019). Syntactic parsing and supervised analysis of Sindhi text. *Journal of King Saud University-Computer and Information Sciences*, *31*(1), 105-112.

Ali, M., & Wagan, A. I. (2017). Sentiment summarization and analysis of Sindhi text. *Int. J. Adv. Comput. Sci. Appl*, *8*(10), 296-300.

Bhurgri A. M. (2010). July 7. A Breakthrough in use of Sindhi on Internet Indus Asia Online Journalhttp://iaoj.wordpress.com/2010/07/07/a-breakthrough-in-use-of-sindhi-on internet/

Collie. J. 2006. The Sindhi Language. In K. Brown (ed). Encyclopedia of Language and Linguistics, 2nd Edition 11: 384-386. Oxford: Elsevier.

CRULP. 2010. Sindhi English Dictionary 2010. http://www.crulp.org/sed/ CRULP Parallel Corpus. Urdu, Nepali and English Parallel Corpus 2010. http://crulp.org/software/ling_resources/UrduNepaliEnglishP-arallelCorpus.html.

Jumani, A. K., Memon, M. A., Khoso, F. H., Sanjrani, A. A., & Soomro, S. (2018). Named entity recognition system for Sindhi language. In *Emerging Technologies in Computing: First International Conference, iCETiC 2018, London, UK, August 23–24, 2018, Proceedings 1* (pp. 237-246). Springer International Publishing.

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)* (pp. 1-5). IEEE.

Ali, M., & Wagan, A. I. (2019). An analysis of sindhi annotated corpus using supervised machine learning methods. *Mehran University Research Journal of Engineering & Technology*, *38*(1), 185-196.

Dootio, M. A., & Wagan, A. I. (2018). Unicode-8 based linguistics data set of annotated Sindhi text. *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 intelligent systems*, *28*(2), 55-63.

Hakro, D. N., Talib, A. Z., Bhatti, Z., & Moja, G. N. (2014). A Study of Sindhi Related and Arabic Script Adapted languages Recognition. Sindh University Research Journal (Science Series), Vol. 46, No. 3, pp. 323-334.

Awan, S. A., Abro, Z. H., Jalbani, A. H., Hakro, D. N., & Hameed, M. (2018). Handwritten Sindhi character recognition using neural networks. *Mehran University Research Journal of Engineering & Technology*, *37*(1), 191-196.

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

Alejandro G., Beatriz A. (2013) "Sindhi", The Languages Gulper, http://www.languagesgulper.com/eng/Sindhi.html. Retrieved December 27, 2013. 3.

Christopher, S., (2023) "Sindhi Language”, Encyclopedia Britannica, https://www.britannica.com/topic/Sindhi-language, [Retrieved December 29, 2018]

Das, A., Bandyaopadhyay, S., & Gambäck, B. (2012). The 5w structure for sentiment summarization-visualization-tracking. In *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I 13* (pp. 540-555). Springer Berlin Heidelberg.

Agrawal, S. S., Abhimanue, S. B., Bansal, S., & Mahajan, M. (2014). Statistical Analysis of Multilingual Text Corpus and Development of Language Models. In *LREC* (pp. 2436-2440).

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

Motlani, R. (2016). Developing language technology tools and resources for a resource-poor language: Sindhi. In *Proceedings of the NAACL Student Research Workshop* (pp. 51-58).

Rahman, M. (2015). Towards Sindhi corpus construction. *Towards Sindhi Corpus Construction, Linguistics and Literature Review, 1*(1), 39-48.

Vamshi Krishna, B., Pandey, A.K., Siva Kumar, A.P. (2018). Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic. In: Cognitive Science and Artificial Intelligence. Springer Briefs in Applied Sciences and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-10-6698-6_8

Negi, S., & Buitelaar, P. (2017). Suggestion mining from opinionated text. *Sentiment Analysis in Social Networks*, 129-139.

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological science*, *14*(1), 60-65.

Riaz, K. (2010). Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop* (pp. 126-135).

Paik, J. H. (2013, July). 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* (pp. 343-352).

Samovar Larry, A., & Porter, R. E. (2006). Intercultural communication.