Architecture of Parts of speech Tagger in Sindhi Language

Saira Baby Farooqui[1*], Noor Ahmed Shaikh[2], Samina Rajper[3],

Abstract:**The Sindhi language is intricate and one of the oldest tongue spoken and written in several parts of the world. In this language words segmentation, The short vowel restoration (SV) and parts of Speech (POS) is tagging the generally challenging jobs for its natural language processing (NLP) applications. Furthermore, it's complex by the features. For example soft spaces in lexis, short vowels are compound and complex words are found in Sindhi. For its complexity, Parts of Speech (POS) tagging are challenging job for machine learning. It can help to overcome these ambiguities. Sindhi has eight POS according to the formation of sentence. The (POS) change their nature that human beings easily understand but a computer does not do. To overcome these issues some rules are defined in this model which may help a machine to recognize POS tagging. In POS Architecture has three contrasting phases to resolve the tagging problems in this language and other languages as well. The tokenization is a utensil for NLP for word segmentation (Sentence are break into words, After segmentation of words SVR phase is start for proper vocalization and applied tags for tagging which can help to understand appropriate texts. POS are tagged in the corpus, it removes the uncertainty. The English language has various rules but the Sindhi language identify its functionality by the corpus (vowel, short vowels, space and white spaces). The architecture helps the society to recognize the indigenous structure of language. This architecture also helpful to translator of Sindhi Language, Question Answering, Information Extraction, Machine Text, Summarization, Translation, Sindhi Dictionaries, Information Retrieval and Web Portals.**

Keywords: **Architecture of Parts of speech tagging, short vowel restoration (SVR), Natural language processing (NLP).**

INTRODUCTION:

Sindhi is complex languages due to syntax. It is being spoken in the United States of America, India, Oman, United Arab Emirates, Singapore, United Kingdom, Indonesia, Pakistan, Iran and many other countries of the world. This language belongs to Indo-European, Northwestern and Indo-Iranian. It is able to be written in various styles of writing such as, Devangari, Gurmukhi, Landa and Persian Arabic. The script of language has many problems like rich morphological structure, primary and secondary words, white or soft spaces in words and the same word belongs to many parts of speeches. Therefore, Tokenization method is used in Sindhi corpus to reduce white and soft space. The Tokenization splits words by which the vowel restoration take place and the tagger puts tags over words according to the structure of sentence. Comparing Sindhi language to Arabic, Sindhi has more vowels and consonants. Each word in Sindhi language ends with a vowel. These are some rules which describe the part of speech in the sentence.

اسم +اسم +فعل= جملو
سنا گانو ڳائي ٿي.
طوطو ٻوليون ٻولي ٿو.
ٻار راند کيڏن ٿا.
اسم+فعل= جملو
ساز وڄو.
خوشحالي آئيندي.
ڇوڪرا ايندا.
اسم + ظرف+ اسم+ حرف جر+ اسم +فعل = جملو
مڇي فقط پاڻيءَ ۾ ترندي آهي.
عورت سدائين گهرن ۾ ڪم ڪندي آهي

سنڌيءَ ۾ اسم، ضمير ۽ هر هڪ پارٽس آف اسپيچ کي فنڪشن جي بنياد تي سڃاڻبو آهي. انگلش ۾ انهن جون نشانيون آهن. پر سنڌيءَ ۾ ڪارج يعني ڪم جي بنياد تي خبر پوندي آهي. سنڌيءَ ۾ صفت هر ان اسم کان اڳ ايندي جنهن جي خاصيت بيان ڪندي، چاهي اهو اسم فاعلي حيثيت ۾ هجي توڙي مفعولي حالت ۾ هجي.

مثال:
سهڻي ڇوڪريءَ گانو ڳايو.
ڇوڪريءَ مٺو انب وغيره.
سنڌيءَ ۾ ظرف هميشه فاعل کانپوءِ ايندو آهي.
مثال:
اسان سڀاڻي ڪراچيءَ وينداسين.
هو اندر ويٺا آهن.
ٻار ٻاهر کيڏن ٿا.
توهان خوبصورتيءَ سان ڳايو.
هن تمام ڏکيو امتحان حل ڪيو.

Sindhi language is the most difficult and rich language having 52 alphabets, short vowel, Tokenization and Parts of speech (POS) tagging. These Parts of Speech (POS) may be changed due to change in placement of short vowels. It explains the result of two noun entities that shows the correlation to these are two entities or two people with third noun.

صنم ۽ حسين علي هڪ ڪلاس ۾ نه آهن. هي. هي ٻئي هڪ اسڪول ۾. به نه آهن

صنم ۽ حسين علي these are two noun entities and هي is pronoun

[1-2-3-]Shah Abdul Latif University, Khairpur Sindh
Country : Pakistan
Email: sairafarooqui@sau.edu.pk

indicating both of them.

LITERATURE REVIEW

Sindhi words always end in a vowel. This vocalic ending is optionally marked by diacritics in written text. Diacritics are also used inside words to represent additional vocalic features [1].

The creation of linguistic applications requires the language corpus. For instance, morphological analysis text to speech diacritics restoration [2] and parts of speech labelling [2 The creation of linguistic applications requires the language corpus. For instance, morphological analysis text to speech diacritics restoration [2] and parts of speech labelling [2]. In practically all NLP applications where the initial stage necessitates tokenizing input into words, word segmentation is the principal mandatory task. The difficulty in word segmentation is a problem for several Asian languages, including Urdu. However, unlike other Asian languages, Urdu suffers both space insertion and omission problems in word segmentation [3]. Every language has a lexicon, which is essential for language learners, researchers, and academics [5]. The technique of properly categorising each word in the text according to its syntactic function is known as part of speech (POS) tagging. Each POS tagger must have a tag set and word disambiguation rules as basic components. Parts-of-Speech tagging (POS) is used with a tag assigned with the intention of using speech to interpret the text. Nouns, verbs, pronouns, and determinants are the main types of tags used by the POS[1,7].

TOKENIZATION:

Tokenization is a word segmentation model in NLP application. Tokenization divides text into meaningful tokens [1]. The Chinese dataset uses traditional Chinese characters. The split was predefined by the UD datasets.[2] The Parts of speech (POS) tagging is a general natural language processing procedure which refers to the classification of words in a group of text into correspondence by a specific part of speech consisted upon definition of word and its context [3][4].

Parts of Speech tags describe the characteristic structure of lexical words within sentence or text; therefore, they are used to make assumptions about semantics. Other applications of parts of speech (POS) tagging including Co-reference resolution, speech recognition, named entity recognition.

مثال:

رات جو آسمان تي تارا هوندا آهن.

آهن	هوندا	تارا	تي	آسمان	جو	رات

Table 1

Tokenization without short vowel

The symbols of letters are called short vowels. These are zair (زير), zabar(زبر), paish (پيش) etc. For SVR, N gram method is used.

مثال:

رَات جَو آسمَانَ تِي تَارَا هُوندَا آهِن.

آهِن	هُوندَا	تَارَا	تِي	آسمَانَ	جَو	رَات

Table 2

Tokenization with short vowel

Taggers are divided into three groups.

Rule base taggers, stochastic taggers, and transformative taggers.

Parts of Speech tagging are used to specify tag and tags are used to identify the words and their relative parts of speech in the corpus. Architecture names the unified parts of speech tagging and its application to Greek language

It has three main phases to achieve target.

1. *Tokenization, short vowel restoration and parts of speech tagging.*

The tokenization is a technique useful for facts and data either "non-case sensitive or case sensitive"[5]. It can also be segment data and identifier with the help of programming kit that facts or data can map to back data through tokenizer [4] [6]. Hence, Sindhi is complicated language for computers for it has two kinds of words.

Primary words ابتدائي يا بنيادي لفظ

Secondary words ثانوي لفظ

Primary words ابتدائي يا بنيادي لفظ have single structure.

For Example: اهو ، پار، هر، رستو. وغيره

Secondary words ثانوي لفظ

The secondary words have more than one shape. The secondary words are further divided into three types:-

Complex words مرتب لفظ

Words having more than one forms are called complex words.

For Example: اڻ سونهون ، پائيتو ، مائيتو ، آپاڳ ،اَجهاڳ ،اڻ واقفيت وغيره

Compound words مرڪب لفظ

When two or more words are combined they will create compound word.

For Example: ڪٺومنٺو، سنئون سڌو،سنهو ٿلهو، وغيره

Reduplicated words دهرايل لفظ يا ٻٽا لفظ

The word repeated two times in a word is called reduplication words.

For Example:زخم زخم،اچ وچ، ڪاڏو پيتو، نفعو نقصان، ، قدم قدم،ِ
ٻڙٻڙ، جيئڻ مرڻ، دوست- دشمن ۽ گهرِ گهر

In Sindhi language most of the alphabets are taken from Persian and mainly of the alphabets are taken form Arabic language and some alphabets are modified letters [7]. These alphabets contain short vowels, they can alter the articulation as well as the POS of the related word according to text [8].

For Example: ان can be أُنَ، اَنُ both words has different meaning and parts of speech. First word أُنَ is pronoun and اَنُ is noun due to occurrences of short vowels.

- سر can be سَرَ، سُرَ ، سُرُ these three sound are dissimilar and its POS and meaning is also diverse. Determine which references in a speech correspond to the same real-world thing, quality, or situation. Co-reference goals is the errand of discovering all thing phrases (NPs) that allude to a similar true substance.

METHODOLOGY

This is architecture of parts of speech tagging. User can enter the words. After the entry of corpus, the tokenizer is activated. Tokenizer is words segmented model which is activated after entry of words. It divides the text into tokens. N gram technique is used for short vowel restoration. These words or tokens put into short vowel restoration (SVR) where N- gram model is trigger. The short vowels tables are given as:-

a. TblUni: TblUni contains single token.
b. TblBi: TblBihave two tokens.
c. TblTri:TblTri contains three tokens.
d. TblQuad: TblQuad have contains four token without short vowels.
e. TblUniV: TblUniVsingle token of word with short vowels.
f. TblBiV: TblBiV consists of two tokens with short vowels.
g. TblTriV: TblTriV having three tokens with short vowel.
h. TblQuadV: TblQuadV have four tokens with short vowels signs.
i. Tbl POS: It contains word with POS or some specific tag.

Noun=N	=Pronoun Pro	=Verb V	=Adverb Ad

Table 3

Adjective=Adj	Preposition=Pre	Conjunction=Con	Interjection=Int

Table4

j. TblMPOS: It contains words with multiple parts of speech. If any ambiguity in the sentence control transfer to the ambiguity phase where ambiguity is resolved. After the step of tokenization all the text will be divided into tokens and occurrence of short vowel restoration over the text. These words are access from sequential storage. If there is no ambiguity, parts of speech will be assigned. If there is ambiguity then resolve the ambiguity and assign the tags with the help of MPOS tagging. POS tagging has two databases one is POS lexicon and second is word lexicon. POS lexicon has IDs, its parts of speeches along with word, lexicon have relational IDs and all the words of corpus. This is a proposed architecture to overcome the challenges of tokenization, SVR and POS tagging.

Architecture of POS Tagger in Sindhi Language

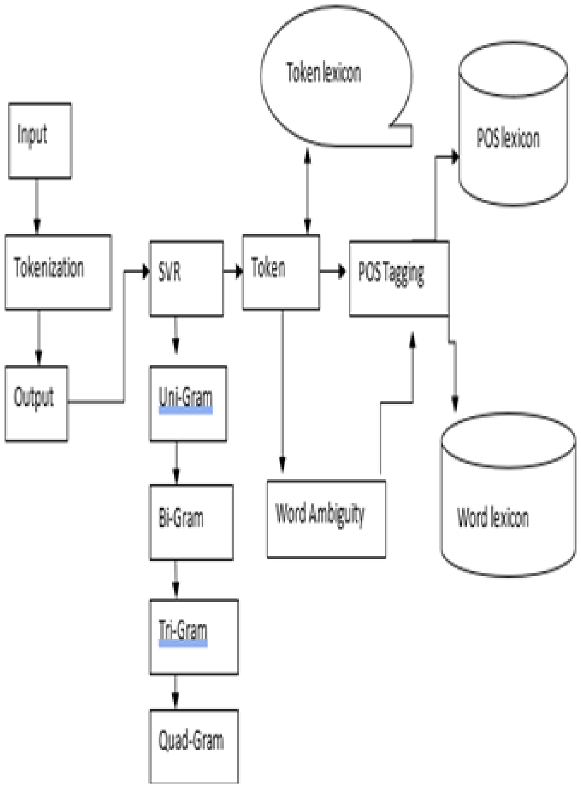
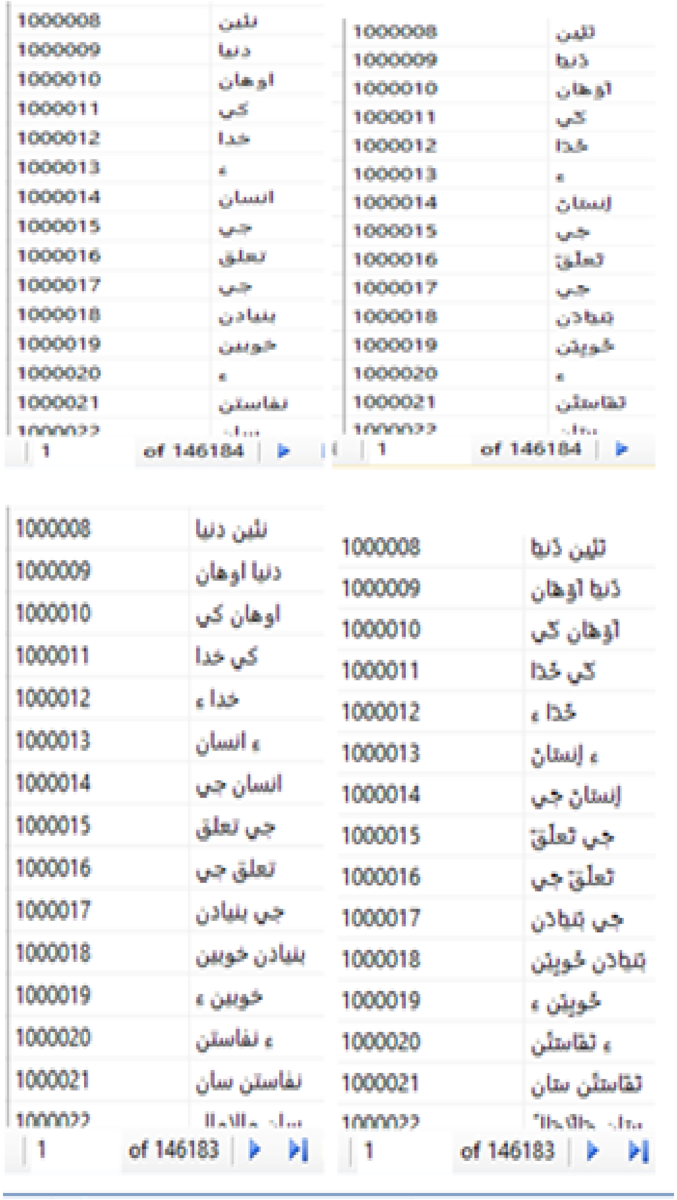


Figure 1



1000008	تلين
1000009	دنيا
1000010	اوهان
1000011	كي
1000012	خدا
1000013	ء
1000014	انسان
1000015	جي
1000016	تعلق
1000017	جي
1000018	بنيادن
1000019	خوبين
1000020	ء
1000021	نفاستن
1000022	[illegible]

1 of 146184

1000008	تلين
1000009	ذنيا
1000010	آوهان
1000011	كي
1000012	خدا
1000013	ء
1000014	إنسان
1000015	جي
1000016	تعلق
1000017	جي
1000018	تنبادن
1000019	خويتن
1000020	ء
1000021	تقاستن
1000022	[illegible]

1 of 146184

1000008	تلين دنيا
1000009	دنيا اوهان
1000010	اوهان كي
1000011	كي خدا
1000012	خدا ء
1000013	ء انسان
1000014	انسان جي
1000015	جي تعلق
1000016	تعلق جي
1000017	جي بنيادن
1000018	بنيادن خوبين
1000019	خوبين ء
1000020	ء نفاستن
1000021	نفاستن سان
1000022	[illegible]

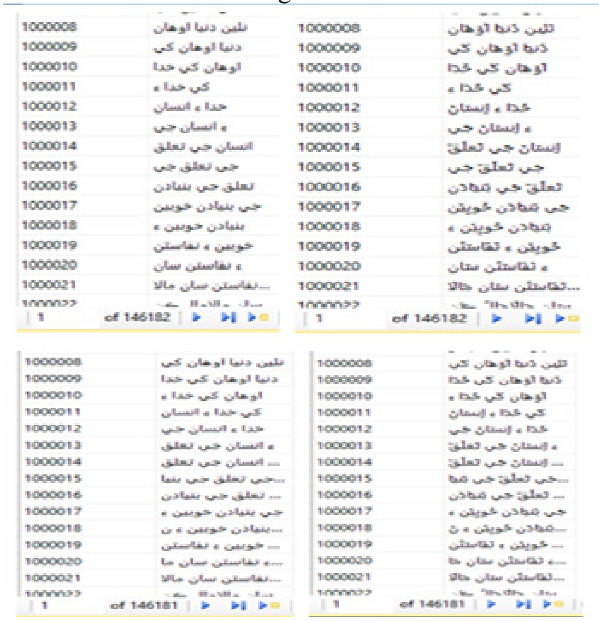
1 of 146183

1000008	تلين ذنيا
1000009	ذنيا آوهان
1000010	آوهان كي
1000011	كي خدا
1000012	خدا ء
1000013	ء إنسان
1000014	إنسان جي
1000015	جي تعلق
1000016	تعلق جي
1000017	جي تنبادن
1000018	تنبادن خويتن
1000019	خويتن ء
1000020	ء تقاستن
1000021	تقاستن ستان
1000022	[illegible]

1 of 146183

Figure 2

Figure 3



1000008	تلين دنيا اوهان
1000009	دنيا اوهان كي
1000010	اوهان كي خدا
1000011	كي خدا ء
1000012	خدا ء انسان
1000013	ء انسان جي
1000014	انسان جي تعلق
1000015	جي تعلق جي
1000016	تعلق جي بنيادن
1000017	جي بنيادن خوبين
1000018	بنيادن خوبين ء
1000019	خوبين ء نفاستن
1000020	ء نفاستن سان
1000021	نفاستن سان مالا...
1000022	[illegible]

1 of 146182

1000008	تلين ذنيا آوهان
1000009	ذنيا آوهان كي
1000010	آوهان كي خدا
1000011	كي خدا ء
1000012	خدا ء إنسان
1000013	ء إنسان جي
1000014	إنسان جي تعلق
1000015	جي تعلق جي
1000016	تعلق جي تنبادن
1000017	جي تنبادن خويتن
1000018	تنبادن خويتن ء
1000019	خويتن ء تقاستن
1000020	ء تقاستن ستان
1000021	تقاستن ستان حالا...
1000022	[illegible]

1 of 146182

1000008	تلين دنيا اوهان كي
1000009	دنيا اوهان كي خدا
1000010	اوهان كي خدا ء
1000011	كي خدا ء انسان
1000012	خدا ء انسان جي
1000013	ء انسان جي تعلق
1000014	انسان جي تعلق ...
1000015	جي تعلق جي بنيا...
1000016	تعلق جي بنيادن ...
1000017	جي بنيادن خوبين ء
1000018	بنيادن خوبين ء ن...
1000019	خوبين ء نفاستن ...
1000020	ء نفاستن سان ما...
1000021	نفاستن سان مالا...
1000022	[illegible]

1 of 146181

1000008	تلين ذنيا آوهان كي
1000009	ذنيا آوهان كي خدا
1000010	آوهان كي خدا ء
1000011	كي خدا ء إنسان
1000012	خدا ء إنسان جي
1000013	ء إنسان جي تعلق
1000014	إنسان جي تعلق ...
1000015	جي تعلق جي تنبا...
1000016	تعلق جي تنبادن ...
1000017	جي تنبادن خويتن ء
1000018	تنبادن خويتن ء ن...
1000019	خويتن ء تقاستن ...
1000020	ء تقاستن ستان حا...
1000021	تقاستن ستان حالا...
1000022	[illegible]

1 of 146181

For the implementation of SVR SQl server is used to create the database.

DISCUSSION:

Almost, researchers believed that Sindhi is complex and old language [9] [10], hence some thought it is complex due to its grammatical structure, although most of them assumed due to its soft space. Some of the researchers told it contains space insertion and omission. Therefore, all are right because it is the mixture of complex, compound and reduplication words, so there may be some issues which may be occur in POS tagging in Sindhi language[11].

Sindhi is believed to be tough language by a number of scholars considering its 52 letters, but some others opinion is that it is no doubt difficult but owing to its writing format.

The alphabets for Persian, Arabic, Urdu, and Sindhi are the same. They employ graphs and orthography. Single graphs show that if a short vowel is absent, numerous different sounds or pronunciations can be produced. Short vowels on alphabets or words may be employed for improved accuracy and comprehension level to eliminate these kinds of problems.[1,4,7]

Many other languages like Sindhi have “starting, middle and ending” structures of the words with starting and ending structures. It contains short vowels on letters, it is believed that it has many styles or laja (لهجو) of speaking and some computer scholars mentioned, it is complex for machine learning. Some people come up with certain questions, that are highly complex language and creates difficulty when learning through machine?

It has primary and secondary classes of words. Sindhi language carries space inclusion mistakes and space exclusion errors. Approximately different languages suppose Urdu, Sindhi Arabic have their entirely different patterns of beginning and ending. This sort of matter creates hurdles for tokenizer. How can it be tokenized?

A number of scholars have proposed tokenizers for different languages. The procedure of tokenization is being used for words segmentation of text into full tokens connotation or meaning just like J Mahar and Bhatti, z. This is proposed architecture of POS tagging. Tokenizer is used for word segmentation. POS tagsets are major prerequisite of POS tagging process in this model. N-gram method is applied for SVR[15]. SVR helps to comprehend the word (Token) and its context. Understanding text and its context the entire token is assigned the tags or parts of speech [16,17]

CONCLUSION:

Sindhi is the most difficult language to identify its POS tagging. Each word of this language has a number of articulation and many POS because it's short vowel depiction (diacritic marks).

Cataloging has a significant application to the natural language processing (NLP) [12]. The POS tagging is considered as an organized segment of labeling syntactic classification to all words [13 [14]. The syntactic classification covers almost all word classes. Example: Noun (اسم), Adjective (صفت), Verb (فعل) and Pronoun (ضمير).

Essentially, POS tagging has been termed as the core portion of NLP functions such as how to identify speech connection between content and words, getting information along with the context and sense of the particular concept [15][18]. Moreover, there are a number of taggers which could be categorized as supervised and un-supervised. In the background of this framework, it is highly necessary to describe the factors of supervised taggers. Those are not able to depend on the post-tagged text but on pre-tagged one. Apart from the above statement, automation is applied for tagging words in the un-supervised category. With words class tagging, rule base approaches are present and have 96% accuracy rate

. FUTURE WORK

Using rule-based techniques, Sindhi rarely sees work on tokenizer, short vowel restoration, or parts of speech labelling. This is a small task that will aid in the creation of word classes and associated tags in the future. Although the focus of this work is mostly on parts of speech tagging, subclasses of these tags and classes will be created in the future. It will make it

REFERENCES

[1] M. S. M. H. &. L. Jacobsen, " Optimal size-performance tradeoffs: Weighing pos tagger models," vol. arXiv preprint arXiv:2104.07951., (2021).

[2] H. J. L. H. W. Yang, "Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector," IEEE Access, Vols. 6, 30174-30183, (2018).

[3] A. H. Aliwy, "Arabic morphosyntactic raw text part of speech tagging system.," (2013).

[4] I. N. A. H. J. a. M. I. C. Sodhar, "Identification of issues and challenges in romanized Sindhi text.," Editorial Preface From the Desk of Managing Editor 10, vol. no. 9), (2019.

[5] P. P. G. V. P. S. Papageorgiou H, "A Unified POS Tagging Architecture and its Application to Greek.," vol. InLREC., 2000 May.

[6] M. L. &. M. U. Rahman, "Towards Transliteration between Sindhi Scripts Using Roman Script, Linguistics and Literature Review," Vols. 1(2), 95-, (2015).

[7] J. A. &. M. G. Q. Mahar, "Sindhi part of speech tagging system using wordnet," International Journal of Computer Theory and Engineering, Vols. 2(4), 53, (2010).

[8] J. A. M. G. Q. &. S. H. Mahar, "Sindhi diacritics restoration by letter level learning approach," Sindh University Research Journal-SURJ (Science Series), vol. 43(2)., (2011).

[9] M. U Rahman, "Towards Sindhi Corpus Construction, Linguistics and Literature Review," vol. 1(1), pp. 39-48, (2015).

[10] H. M. J. A. &. M. M. H. Shaikh, "Statistical approaches to instant diacritics restoration for sindhi accent prediction," Sindh University Research Journal-SURJ (Science Series), vol. 49(2)., (2017).

[11] J. J. &. K. C. Webster, "Tokenization as the initial phase in NLP," The 14th International Conference on Computational Linguistics., vol. Volume 4, 1992 Volume 4:.

[12] M. Attia., "Arabic Tokenization_System," Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, vol.

https://www.researchgate.net/publication/234810209, p. 65–72, Date accessed: 01/2007 .

[13] L. F. Schmid H, “Estimation of conditional probabilities with decision trees and an Application to fine-grained pos tagging.,” vol. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), p. 777–784, August (2008).

[14] J. A. &. M. G. Q. (. F. Mahar, “. Rule based part of speech tagging of sindhi language,” International Conference on Signal Acquisition and Processing, vol. IEEE, pp. (pp. 101-106), February (2010).

[15] A. Hardie, “ Developing a tag set for automated part-of-speech tagging in Urdu.,” vol. In Corpus Linguistics 2003., (2003).

[16] Y. Hifny, Hifny, Y. (2012). Smoothing techniques for Arabic diacritics restoration. In Proceedings of the 12th Conference Lang. Eng.(ESOLEC’12) (No. 1, pp. 6-, no. 1, pp. (No. 1, pp. 6-12).

[17] I. S. J. A. S. R. ZITOUNI, “Maximum entropy-based restoration of Arabic diacritics,” In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics , no. Association for Computational Linguistics., pp. (pp. 577-584)., 2006, July.

[18] K. R. B. S. P. a. K. D. Singha, “”Part of Speech Tagging in Manipuri: A Rule based Approach,” International Journal of Computer, p. 51, (2012) .