Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2768

2024, 43(3) 47-55

Parts-of-speech tagger for Sindhi language using deep neural network architecture

Adnan Ali Memon [a], Saman Hina [b], Abdul Karim Kazi [b,*], Saad Ahmed [c]

[a] *Department of Software Engineering, BUITEMS University, Quetta, Pakistan*

[b] *Department of Computer Science and Information Technology, NED University of Engineering and Technology, Karachi, Pakistan*

[c] *Department of Computer Science, Iqra University, Karachi, Pakistan*

[*] Corresponding Author: Abdul Karim Kazi, Email: karimkazi@neduet.edu.pk

Received: 01 March 2023, Accepted: 14 June 2024, Published: 01 July 2024

KEYWORDS

Sindhi Parts of Speech

Parts of Speech (POS) Tagging

Sindhi Corpus

Long-Short Term Memory (LSTM)

Gated Recurrent Unit (GRU)

ABSTRACT

Language is a fundamental medium for human communication, encompassing spoken and written forms, each governed by grammatical rules. Sindhi, one of the oldest languages, is characterized by its rich morphology and grammatical structure. Part-of-speech (POS) tagging, a crucial process in natural language processing, involves assigning grammatical tags to words. This research presents a novel approach to POS tagging for Sindhi text using deep learning techniques. We developed a POS tagger employing Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, with LSTM demonstrating superior effectiveness. This study represents the first application of these deep learning methods for POS tagging in Sindhi. Utilizing fastText, we trained 79,959 Sindhi word vectors, derived from a corpus compiled from diverse sources including Sindhi books, stories, and poetry. The corpus comprises 1,459 sentences and 10,584 unique words, split into 80% for training and 20% for validation. Our results indicate that the LSTM model achieved an accuracy of 85.80%, outperforming the GRU model, which achieved 80.77%, by a margin of 5%. This work's novelty lies in the application of deep learning techniques to enhance POS tagging accuracy in the Sindhi language corpus.

1. Introduction

The Sindhi language is part of the Indo-Aryan language spoken by Pakistani and Indian people. In Pakistan, it is written in a slightly different form of the Perso Arabic script, with additional letters to accommodate implosive, retroflex, and nasal sounds [2]. Sindhi is written from right to left with 52 alphabetic characters, as shown in Fig. 1. Sindhi language origins can be traced back to 1500 BC [3, 4]. The name ‘Sindhi’ is derived from the name of a river, known as the “Indus River” or “Sindhu”. The Sindhi language is also registered as the official language of two countries: Pakistan and India. With the world becoming a global centre, people have access to a plethora of information that may be utilised for both internal and cross-cultural communication as well as engagement across languages and civilizations. Sindhi language communication is increasing every day. In addition to this, advancements in technology are helping people understand languages in a better manner. Nowadays, there is an abundance of interest in research on natural languages. Therefore, it is becoming more important to incorporate new techniques.

Natural language processing (NLP) is one of the important fields of artificial intelligence, which is the process of developing software applications that enable computers to understand natural languages like English, Urdu, Sindhi, Arabic, German, Hindi, Chinese, and many others. A substantial amount of research exists, spanning various languages [34-38]. Some languages are challenging for many linguistic tasks such as Arabic [39] and Sindhi. Some known applications that are widely used for these languages include “Parts of Speech (POS) tagging”, “Named Entity Recognition (NER)” [4], “Machine Translation (MT)”, “Information retrieval (IR)”, “Information extraction (IE)”, “Morphological Analysis”, “Syntactic parsing” [5], etc. However, researchers find it challenging to work on NLP tasks using local languages. This is due to the lack of availability of a corpus of Sindhi language [6]. We know that natural language is available digitally in different formats (such as audio, images, etc.) but, this research focuses on the POS tagging of a corpus containing text in the Sindhi language. Parts-of-speech Tagging is a process of grammatically marking words or assigning the appropriate lexical category to words in sentences in any natural language. Some categories are nouns, pronouns, adjectives, verbs, adverbs, prepositions, and conjunctions. POS tagging is an important sub-component or a pre-requisite process of human natural language processing tasks. This is mostly used in machine learning for translation from the source language to the desired language, named entity recognition, spell checker, syntactic parsing, information retrieval, etc.

The Sindhi language is the official language of the Sindh province in Pakistan. A lot of work has already been done on parts of speech tagging in different languages such as English [7], Urdu, Hindi, etc., but very little work has been reported on the Sindhi parts of speech tagging. There is a lot of room available for research on the Sindhi parts of speech tagging by using different techniques. In the reported studies, researchers have used support vector machines for empirical and statistical studies on the Sindhi language [8] but no one has worked on POS tagging using long short-term memory (LSTM) and gated recurrent unit (GRU). This combination of techniques (LSTM and GRU) is not only the suitable algorithm in general for computers to learn natural language but also has proved to be effective for Sindhi POS taggers using word embedding.

Another problem for a computer is finding out or understanding the ambiguity [9] of natural language, which is a challenging problem in the case of Sindhi. Humans can easily understand natural languages, but, on the other hand, computers cannot understand the particular context of a word in any given sentence of any natural language. This is the reason why parts of speech tagging are used as linguistic features to help computers understand the context and ambiguity in natural languages. In this process, more and more tagged information is provided to the computers, and then using ML techniques, computers can learn and understand human natural language in a better way for different applications [10-11]. Generally, three approaches may be adopted for the POS tagging problem: the first rule-based approach [12], the second corpus-based approach [13], and the third hybrid approach [14–15, 33]. Rule-based POS tagging uses a set of manually developed dictionaries to tag words in given sentences. Developing rule-based POS tagging for the Sindhi language is very hard because the Sindhi language is morphologically vibrant. A POS tagger that learns from a corpus using a pre-annotated Sindhi language corpus is proposed to label or tag the unlabelled data accurately.

2. Related Work

In contrast with the literature related to POS tagging, a limited number of studies have been conducted on parts of speech tagging for the Sindhi language. This section contains details regarding previous research in which researchers have reported the use of different machine learning techniques for POS tagging.

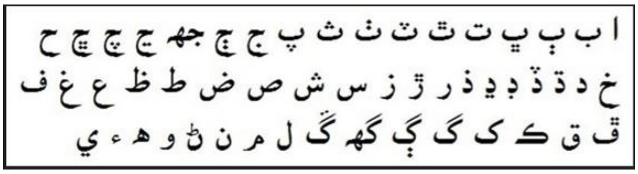

Fig. 1. Sindhi 52 Alphabets [3]

A POS tagger for the Sindhi language was integrated into a Sindhi text processing system created by a team of researchers [16]. These researchers have also talked about how to explain various contexts and get rid of ambiguity that can arise in different portions of speech. Because of the word morphology, Sindhi POS tagging is more difficult; they have discussed several POS approaches. A corpus of 5,000 words was used for training and 2000 words for testing in the Sindhi language. Experts in the Sindhi language manually tagged these words. The authors [6] created an annotated corpus of the Sindhi language using both the Sindhi Parts of Speech tag sets and the Universal Parts of Speech tag set. Term frequency and inverse document frequency have been employed by these scholars. Furthermore, they developed a supervised machine learning model to grammatically evaluate the Sindhi annotated corpus. The model was trained on 80% of the dataset and tested on the remaining 20%. For validation, they employed a 10-fold cross-validation technique. The SVM Non-Linear model achieved an accuracy of 89.16% using Universal Part-of-Speech (UPOS) tags and 89.1% using Sindhi Part-of-Speech (SPOS) tags. In contrast, the Random Forest method attained significantly higher accuracies of 99.57% with UPOS and 99.89% with SPOS.

In another research, the authors reported the rule-based approach for semantic Sindhi parts of speech tagging [2]. This approach relies on a WordNet lexical database to identify the relationships between words in a particular text. Moreover, these researchers described the Sindhi POS tag set and also worked on word-sense disambiguation algorithms that were developed and designed for POS tagging. In their research, they have used two types of lexicons: one for simple words and the other for disambiguated words. The corpus is collected from the Sindhi Dictionary, and the developed model was tested on the Sindhi word lexicon (SWL) [31] that was developed by these researchers and the WordNet lexicon (WNL) [2]. The SWL contains '26366' tagged words. The WNL lexicon contains 1885 analogical words. The accuracy of 96.28% was achieved without the use of WordNet. Similarly, with the WordNet approach, the accuracy increased to 97.14%. The author also observed that when poetry and future words were used, accuracy became low.

Another reputable work on Sindhi POS tagging was conducted using a machine learning approach [17]. The authors used a machine learning approach named Support Vector Machine (SVM) to tag the sentence with Sindhi POS tags. These researchers collected a corpus of a corpus of 28000 words from different internet resources (poetry, primary school textbooks, newspapers, and stories), and they used 67 tags. The authors reported a good accuracy of 97.86% as compared to their previous work [18]. Another development in the same research domain was presented by [19]. The author applied Sindhi Unicode-8-based data, which is a multiclass and multi-featured dataset. This dataset shows information on the grammatical and morphological structure of Sindhi language text. According to the author, this data will be useful for information retrieval, semantic analysis, and sentiment analysis of the Sindhi language. The Sindhi corpus is processed for annotation and sentiment analysis in the author's tool for the Sindhi NLP application (https://sindhinlp.com). The Sindhi corpus was processed to perform sentiment analysis and annotation in the Sindhi NLP tool separately. The unigram model is used to calculate the probability of every lexicon that is present in the Sindhi corpus. The Farther dataset is processed for normalization and statistical analysis. The same researcher pointed out the problem in the development of Sindhi text corpora due to the lack of resources for computational data [6]. The author first collected data from different online resources, such as books, newspapers, magazines, blogs, and other online websites. All these resources were utilized to build a Sindhi text corpus. Then the authors adopted Document-Term Matrix DTM and TF-IDF techniques and applied them to the analysis performed using the n-gram model. These researchers used a supervised model to formulate it by using SVMs and KNN techniques to perform analysis on the Sindhi sentiment analysis corpus dataset. Precision, recall, and f-score show better performance. Cross-validation techniques are used with 10 folds to validate and evaluate data sets randomly for supervised machine learning analysis. In Sindhi NLP, another study has been carried out to summarise the existing work on Sindhi Language Processing (SLP) and highlight the importance of the Sindhi language [20]. This study emphasized the challenges of the Sindhi language in terms of its computational processing, morphological characteristics, and structure. The research was useful to explore potential NLP applications in the Sindhi language. This paper will be helpful for the researchers to find all the information regarding SLP in one place in a unique way. As a result, important applications include part-of-speech (POS) taggers, spell checkers, diacritic restoration systems, Text-to-Speech (TTS) synthesis systems, Optical Character Recognition (OCR), and Machine Translation (MT) systems. The corpus of the language is necessary for the development of the linguistic applications of either Sindhi or another human language, for instance, parts of speech tagging [21].

According to the research study of Sindhi text [22], the author gives a concept of a model for segmenting Sindhi text into a word tokenization. The author downloaded the Sindhi corpus from different internet resources. The main task for the author is to segment the Sindhi words into word tokens. He faced difficulty in finding the correct word segmentation. To solve this problem, the author used three different layers. The model consists of three layers: Layer One is used to input the text and segment the words using white space; simple and compound words are segmented in Layer Two; and complex words are segmented in Layer Three. It achieved an accuracy of 91.76%. The tokenizer is tested on 2792 Sindhi words.

In contrast with the research work done for the Sindhi language, the presented research focuses on the enhancement of the existing corpus (that includes '10584' distinct words) and the POS tagging using deep learning approaches (LSTM and GRU) that have not been explored for the Sindhi language.

3. Annotation and Collection of Corpus

The corpus used in this research is a combination of the available Sindhi corpus tagged with Universal POS and Sindhi POS tag sets [18], along with enhancements from different resources. These resources include input from a domain expert, internet resources, and handwritten text extraction from Sindhi history books. To develop the gold standard version of this enhanced corpus, researchers manually read Sindhi grammar and also learned it from primary teacher (teaching Sindhi) to annotate the POS tags without errors. Two domain experts from different levels (academic and native speakers) were involved in the annotation and validation process. After the validation by domain experts, the gold standard corpus was completed to be used in this research. The developed models in this research were used on both the original corpus [3] and the enhanced corpus for the analysis. The sample of the enhanced corpus that was designed is presented in Fig. 2.

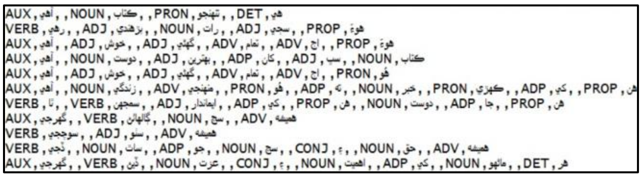

Fig. 2. Sample of Sindhi Annotated Corpus

The complete corpus was annotated according to the explained format and used a new line for the next sentence. The details of the annotated corpus are presented in Table 1.

Table 1

Details of corpus

Details	Count
Total number of words used in the corpus	17312
Total number of sentences used in the corpus	1959
Total number of distinct words used in the corpus	10584
Maximum sentence size of words in the corpus	35

As it is clear that the Sindhi annotated corpus is essential for the Sindhi NLP application, it is important to develop the POS tagger as an important pre-processing resource for the Sindhi language [23]. The development of an annotated corpus was one of mentioning tags just after the token, separated by a comma. Based on the complete analysis of the Sindhi corpus, we were able to identify the number of tags for each Sindhi word. We have used Universal Parts of Speech (UPOS) tagging to tag our Sindhi corpus. It is important for the POS tagger and POS-tagged corpus of any language to define their tag set. Different types of tag sets are available for use in natural language processing, Universal POS is one of them, and our Sindhi annotated corpus is tagged with the UPOS tag set, which is very useful for annotating the Sindhi corpus. Table 2 shows the tag types and the count for this Sindhi corpus.

Table 2

Details of tags distribution in the corpus

UPOS Tag	Tag Type Descriptions	Count
NOUN	Nouns	2961
PROPN	Proper Nouns	1875
PRON	Pronouns	2919
DET	Determiners	1502
VERB	Verbs	3168
AUX	Auxiliary verbs	787
ADJ	Adjectives	927
ADV	Adverbs	894
ADP	Ad position	1220
CONJ	Conjunctions	883
NUM	Number	164
X	Unknowns	16

Corpus linguistics is the analysis of naturally occurring languages based on computerised corpora. Usually, the analysis is performed with the help of a computer, i.e., with specialised software, and takes into account the frequency of the phenomena investigated. The Sindhi corpus used in this research was collected from different sources (like Sindhi books, poems, and stories) [24]. Data can be collected from different sources as well, like newspapers, blogs, and more [25]. This collection of data also had non-Sindhi words that were manually removed. Manual pre-processing is time-consuming but effective in terms of validation by domain experts. However, verbal agreement on data has been reported by the domain experts; therefore, no validation scores can be calculated. Sample data that was collected and cleaned is presented in Fig. 3.

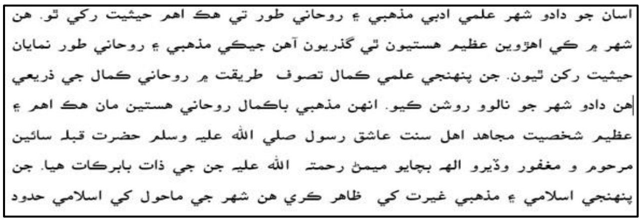

Fig. 3. Sample Data After Removing Non-Sindhi Words

4. Sindhi POS Tagging Using Deep Learning Model

After applying LSTM and GRU techniques to find the accuracy of each part of speech, different accuracies have been observed, tabulated in Table 3.

Table 3

Accuracy of each POS tags

POS	GRU %	LSTM %
NOUN	83.6	89.2
PROPN	71.0	83.9
PRON	76.8	72.7
DET	90.1	91.4
VERB	85.3	88.0

AUX	68.2	72.4
ADJ	83.8	84.3
ADV	88.6	90.1
ADP	60.3	67.8
CONJ	61.6	69.2
NUM	93.8	95.1

4.1 Cleaning of Corpus

As previously mentioned, non-Sindhi that were irrelevant were removed from the corpus by domain experts. Additionally, misspelt Sindhi words were manually corrected with the assistance of domain experts. These experts, including an academic (primary school teacher) and a native speaker, thoroughly validated the data, ensuring the removal of unnecessary words from the corpus, which comprised 17,312 words and 1,959 sentences. We undertook several steps to clean the corpus and eliminate data that was unsuitable for the application of the deep learning model. The cleaning steps are as follows:

- Removing Punctuation: The corpus contained numerous punctuation marks, which were unnecessary for this research. We used a simple function from NLTK to remove these punctuations from the Sindhi corpus.

- Removing Stop Words: After removing punctuation, stop words were eliminated to further cleanse the corpus of irrelevant information. Stop words do not aid in identifying particular POS tags. To remove these stop words from our Sindhi dataset, we created a specialized function.

4.2 Tokenization

Tokenization is the process of breaking the text into small chunks or words. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. For tokenization, NLTK, Genism, and Keras libraries were used in this research. The process of separating or segmenting this Sindhi input sequence of symbols into a particular token known as tokenization [22], is shown in Fig. 4.

```
['آهي', 'پيار', 'سان', 'سنڌ', 'کي', 'اسان']
['AUX', 'NOUN', 'ADP', 'PROPN', 'ADP', 'PRON']
```

Fig. 4. Sindhi Word Tokenization

After tokenization and sentence identification, corpus was divided into training and test data (as shown in Fi 5). It has been observed that the proposed model(s) using LSTM and GRU models have produced better performance which was not employed previously for Sindhi POS tagging.

The parameters of the LSTM neural network used by the POS tagger for the accurate prediction using validation datasets. The LSTM neural network was suggested for clarification of series and textual data related problems [26]. We experimented on our Sindhi corpus with the LSTM model. LSTM uses a weighted sum of previous inputs at each neuron with a nonlinear unit.

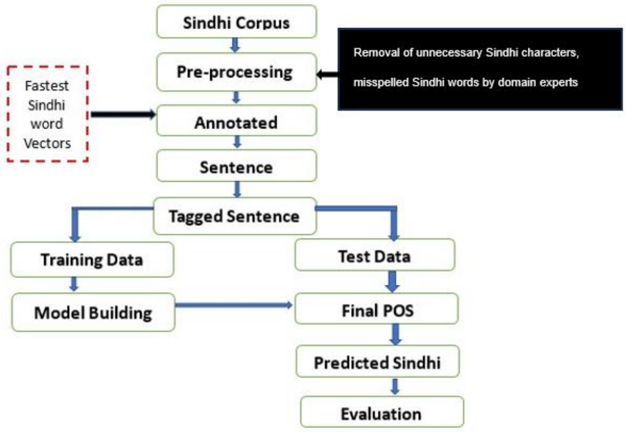

Fig. 5. The Architecture of Sindhi POS Tagger

The LSTM model is divided into three layers. The embedded layer, or input layer, is the first layer of the model embedding layer that uses one hot vector to define the representation of several inputs with a particular size of input dimension. It was important to separate tagged words (tokens) from their tags for further processing. These tokens were stored in matrix form for both training and testing sentences. Here, the input dimension of the first embedding layer is the size of the word vocabulary, as shown in Fig. 6. Moreover, parts of speech (POS) tagging in the Sindhi language uses a one-hot vector that represents every word in the language based on its grammatical category. Here in Fig. 6, each unique part of speech is assigned an index, and the vector of length equal to the total number of unique POS tags in the Sindhi language is used. A word's vector will be zero at all other indices and one at the index where its POS tag is found. This method allows models to interpret and learn from POS-tagged Sindhi text by converting category data into a numerical representation appropriate for computational techniques.

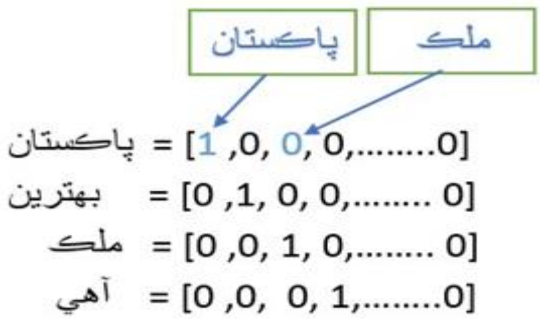

Fig. 6. Sample of One Hot Vector Input Dimension

The second layer of LSTM is to remember the information for a long time; this is its best feature for

NLP. This neural network is explicitly designed to circumvent the long-term dependency in natural language processing problems, where all the LSTM have a particular connection to the repeating module of the neural network. The repeating module or looping module of the LSTM network has four layers that interact in a particular manner. The output layer is the third layer, and this is a softmax layer. The dimension of the output layer is the count of tags or the number of tag types in the given Sindhi corpus. All the training weights use the Stochastic Gradient Descent algorithm [27] to maximize the Sindhi corpus training data.

A gated recurrent unit (GRU) [28] works similarly to LSTM [29]. GRU has two gates or parameters one of them is the reset gate and another one is the update gate. The update gate is the same as LSTM's forget and input gate. The main purpose of the update gate is to identify which information is to be discarded and which information is to be retained and added. At the same time, the reset gate is used to work on past information, it determines how much past information is to be discarded and how much is to be retained. We have used 79959 word vectors for our Sindhi annotated corpus with 300 dimensions. As per our literature review, no one has used the word vectors in Sindhi POS tagging. Our proposed work is a novel approach in Sindhi POS tagging using 300 dimensions. Word vectors using 300 dimensions is shown in Fig. 7.

```
[ ]  from keras.preprocessing.sequence import pad_sequences
     y = pad_sequences(tag, maxlen=max_phrase_len, padding='post')
     print(X[0])
     print(y[0])

     [['0.505752231957471' '-0.8086086184838925' '-0.5556658168419453' ...
       '-0.3295467721766094' '-0.394299002689245' '0.680374356982945']
      ['-0.0107' '-0.0460' '0.0132' ... '-0.0399' '0.0252' '0.0615']
      ['-0.0138' '-0.0484' '-0.0022' ... '-0.0173' '0.0096' '-0.0286']
      ...
      ['0' '0' '0' ... '0' '0' '0']
      ['0' '0' '0' ... '0' '0' '0']
      ['0' '0' '0' ... '0' '0' '0']]
     [ 7  6  5 13  5  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
       0  0  0  0  0  0  0  0  0  0  0]
```

Fig. 7. Tagged Sentence With 300 Dimensions Word Vectors

5. Experiments and Results

For our experiments, the corpus was divided into two sets: 80% (1,268 sentences) for training and 20% (192 sentences) for testing, with words randomly selected from the corpus. Details of the corpus used in this research is provided in Table 4. Moreover, two machine learning techniques: LSTM and GRU were utilized in the experiments. Both techniques are types of 'Recurrent Neural Networks and are good for capturing long-term dependencies. In addition to this, both LSTM and GRU are best for sequential data as compared to SVM. However, another critical aspect is that the LSTM and GRU are working 3 gates and 2 gates respectively, that require memory of previous states which gives better performance. These techniques have not been used by previous researchers for POS tagging of the Sindhi language.

Table 4

Corpus statistics

Details	Count
Total	16312
Total sentences	1459
Total distinct words	10584
Maximum sentence size	35
Total training sentence	1268
Total testing sentence	192

Since accuracy represents the percentage of successfully tagged words, it is a justified and relevant metric for POS tagging since it gives a clear and understandable indication of model performance. The impressive validation accuracies are particularly attributed to the representational power of LSTM [30] tabulated in Table 5. Both models were tested over various epochs (20, 40, 60, 80, and 100) using the Sindhi annotated corpus. It has also been observed that the LSTM model shows a steady increase in both training and validation accuracy as the number of epochs increases. This indicates that the model continues to learn and improve with more training time. On the other hand, GRU model also shows relative improvement with more epochs, but the gains in validation accuracy are slightly less pronounced compared to the LSTM model.

Table 5

Accuracies of models over various epochs

Model	Epochs	Max. validation accuracy	Max. training accuracy
LSTM	20	75.63%	81.10%
	40	79.53%	87.06%
	60	79.49%	90.09%
	80	80.00%	92.88%
	100	80.96%	95.61%
GRU	20	72.35%	80.09%
	40	76.17%	86.50%
	60	77.45%	90.77%
	80	79.76%	93.69%
	100	80.00%	94.50%

Both models show significant increases in training accuracy with more epochs, but validation accuracy plateaus, suggesting overfitting. The LSTM model reaches 95.61% training accuracy, while the GRU model has 94.50% training accuracy and 80.00% validation accuracy, suggesting overfitting.

6. Conclusion and Future Work

Part-of-speech (POS) tagging is a fundamental task in Natural Language Processing (NLP), crucial for developing various applications. For the Sindhi language, POS tagging serves as an essential pre-processing step, labeling each word in the text with its appropriate grammatical tag. This research introduces a novel approach by employing deep learning techniques for POS tagging within the Sindhi corpus. Although deep learning methods have been used for POS tagging in various languages [31], this study specifically explores the efficacy of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for tagging Sindhi text.

In this comparative study, both LSTM and GRU models were evaluated using a manually developed and verified Sindhi gold standard corpus comprising 17,312 words. The annotated corpus was divided into a training set (80%) and a validation set (20%) to assess model performance accurately. Contrary to some literature suggesting that GRU, being a simplified version of LSTM, might perform comparably, our experiments showed that LSTM outperformed GRU by approximately 5%. The results indicate that the LSTM model is better suited for handling the morphological richness and inherent ambiguity of the Sindhi language [24]. The three-gate mechanism of LSTM allows it to manage large datasets more effectively [32], leading to higher accuracy in POS tagging compared to the GRU model. Consequently, the deep learning approach leveraging LSTM has demonstrated superior performance, making it a robust choice for POS tagging in the Sindhi language. In summary, this study highlights the potential of LSTM in enhancing the accuracy of POS tagging for Sindhi, a language characterized by significant morphological complexity. These findings contribute valuable insights to the field of NLP, particularly for languages with similar linguistic challenges

7. References

[1] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, “RuSentiment: An enriched sentiment analysis dataset for social media in russian”, Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018, pp. 755–763. Accessed: Aug. 23, 2021. [Online]. Available: https://aclanthology.org/C18-1064

[2] J. A. Mahar and G. Q. Memon, “Sindhi part of speech tagging system using wordnet”, IJCTE, pp. 538–545, 2010, doi: 10.7763/IJCTE.2010.V2.198.

[3] M. A. Dootio and A. I. Wagan, “Syntactic parsing and supervised analysis of Sindhi text”, Journal of King Saud University - Computer and Information Sciences, vol. 31, no. 1, pp. 105–112, Jan. 2019, doi: 10.1016/j.jksuci.2017.10.004.

[4] Jumani, A. K., Memon, M. A., Khoso, F. H., Sanjrani, A. A., & Soomro, S., "Named entity recognition system for Sindhi language", Proceedings of the First International Conference on Emerging Trends in Computer Engineering and Information Technology (iCETiC 2018), London, UK, August 23–24, 2018, pp. 20-29. DOI: [10.1007/978-3-319-95450-9_20](https://doi.org/10.1007/978-3-319-95450-9_20).

[5] J. A. Mahar and G. Q. Memon, “Rule-based part of speech tagging of sandhi language”, 2010 International Conference on Signal Acquisition and Processing, Feb. 2010, pp. 101–106. doi: 10.1109/ICSAP.2010.27.

[6] M. A. Dootio and A. I. Wagan, “Development of Sindhi text corpus”, Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 4, pp. 468–475, May 2021, doi: 10.1016/j.jksuci.2019.02.002.

[7] M. Dootio, “An analysis and solution of computational linguistics problems of Sindhi text”, Language, vol. 10, pp. 8–24, Dec. 2018.

[8] I. H. Sodhar, D. Jalbani, and A. H. Buller, “An empirical and statistical study on POS tagging of Sindhi social media text”, Feb. 2020. doi: 10.13140/RG.2.2.34057.80487.

[9] S. Mittal, N. Sethi, and S. Sharma, “Part of speech tagging of Punjabi language using n-gram model”, International Journal of Computer Applications, vol. 100, pp. 19–23, Aug. 2014, doi: 10.5120/17634-8229.

[10] M. Imran, S. Hina, and M. M. Baig, “Analysis of learner’s sentiments to evaluate sustainability of online education system during COVID-19 pandemic”, Sustainability, vol. 14, no. 8, 2022, doi: 10.3390/su14084529.

[11] A. Qaiser, S. Hina, A. K. Kazi, R. Asif, and S. Ahmed, “Fake news encoder classifier (FNEC) for online published news related to COVID-19 vaccines”, Intelligent Automation and Soft Computing, vol. 37, no. 1, pp. 73–90, 2023, doi: 10.32604/iasc.2023.036784.

[12] E. Brill, (1992). "A simple rule-based part of speech tagger", Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, pp. 152–155. Association for Computational Linguistics. DOI: [10.3115/974499.974526](https://dl.acm.org/doi/10.3115/974499.974526).

[13] M. Tachbelie, S. Abate, and L. Besacier, "Part-of-speech tagging for under-resourced and morphologically rich languages–the case of Amharic", Human Language Technologies for Development (HTLD), Apr. 2011.

[14] S. Singh, K. Mohnot, N. Bansal, and A. Kumar, "Hybrid approach for part of speech tagger for Hindi language", International Journal of Computer Technology and Electronics Engineering, Feb. 2014.

[15] H. Rana, M. U. Farooq, A. K. Kazi, M. A. Baig, and M. A. Akhtar, "Prediction of agricultural commodity prices using big data framework", Eng. Technol. Amp Appl. Sci. Res., vol. 14, no. 1, pp. 12652–12658, Feb. 2024, doi: 10.48084/etasr.6468.

[16] G. G. Junejo, M. S. H. Talpur, T. Nuzhat, and S. H. Talpur, "POS (parts of speech) tagging system for Sindhi language", International Journal of Computer Science and Emerging Technologies, vol. 4, no. 2, Art. no. 2, 2020.

[17] F. A. Surahio, and J. A. Mahar, "Prediction system for Sindhi parts of speech tags by using support vector machine", Proceedings of 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-6, DOI: 10.1109/ICOMET.2018.8346331.

[18] M. Ali and A. I. Wagan, "An analysis of Sindhi annotated corpus using supervised machine learning methods", Mehran University Research Journal of Engineering and Technology, vol. 38, no. 1, pp. 185–196, Jan. 2019, doi: 10.22581/muet1982.1901.15.

[19] M. A. Dootio and A. I. Wagan, "Unicode-8 based linguistics data set of annotated Sindhi text", Data in Brief, vol. 19, pp. 1504–1514, Aug. 2018, doi: 10.1016/j.dib.2018.05.062.

[20] W. A. Jamro, "Sindhi language processing: A survey", 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), Apr. 2017, pp. 1–8. doi: 10.1109/ICIEECT.2017.7916560.

[21] H. Sajjad and H. Schmid, "tagging Urdu text with parts of speech: A tagger comparison", Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, Mar. 2009, pp. 692–700. Accessed: Oct. 01, 2021. [Online]. Available: https://aclanthology.org/E09-1079

[22] Mahar, H. Shaikh, and G. Memon, "A model for Sindhi text segmentation into word tokens", Sindh University Research Journal (Science Series), vol. Volume 44, pp. 43–48, Mar. 2012.

[23] S. Shah, I. Ismaili, Z. Bhatti, and A. Waqas, "Designing XML tag-based Sindhi language corpus", 2018 International Conference on Computing, Mathematics and Engineering Technologies, Mar. 2018, pp. 1–5. doi: 10.1109/ICOMET.2018.8346381.

[24] I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, "Identification of issues and challenges in Romanised Sindhi text", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 9, Art. no. 9, 36/30 2019, doi: 10.14569/IJACSA.2019.0100929.

[25] M. U. Rahman, "Towards Sindhi Corpus Construction," Linguistics and Literature Review, vol. 1, no. 1, pp. 39-48, Mar. 2015. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3820418.

[26] D. Sachan, M. Zaheer, and R. Salakhutdinov, "Investigating the working of text classifiers", Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018, pp. 2120–2131. Accessed: Aug. 23, 2021. [Online]. Available: https://aclanthology.org/C18-1180

[27] I. A. Kandhro, S. Z. Jumani, A. A. Lashari, S. S. Nangraj, Q. A. Lakhan, and M. T. B. and S. Guriro, "Classification of Sindhi headline news documents based on TF-IDF text analysis scheme", INDJST, vol. 12, no. 33, pp. 1–10, Sep. 2019, doi: 10.17485/ijst/2019/v12i33/146130.

[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", arXiv:1412.3555 [cs], Dec. 2014, Accessed: Oct. 01, 2021. [Online]. Available: http://arxiv.org/abs/1412.3555

[29] R. N. Patel, P. B. Pimpale, and M. Sasikumar, “Recurrent neural network-based part-of-speech tagger for code-mixed social media text”, arXiv:1611.04989 [cs], Nov. 2016, Accessed: Oct. 01, 2021. [Online]. Available: http://arxiv.org/abs/1611.04989

[30] Z. Teng and Y. Zhang, “Two local models for neural constituent parsing”, Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018, pp. 119–132. Accessed: Aug. 23, 2021. [Online]. Available: https://aclanthology.org/C18-1011

[31] G. Prabha, P. Jyothsna, K. Shahina, P. B., and S. Kp, “A deep learning approach for part-of-speech tagging in Nepali language”, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Sep. 2018, pp. 1132–1136. doi: 10.1109/ICACCI.2018.8554812.

[32] S. A. Chowdhury and R. Zamparelli, “RNN simulations of grammaticality judgments on long-distance dependencies”, Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018, pp. 133–144. Accessed: Aug. 23, 2023. [Online]. Available: https://aclanthology.org/C18-1012.

[33] A. K. Kazi, M. Andleeb, S. Ahmed, R. Asif, and Nabeel, “Predictive healthcare analysis of Pakistan’s COVID-19 pandemic using data mining and time series modelling”, PJETS, vol. 11, no. 1, pp. 74-84, Jan. 2024.

[34] B. Basumatary, M. Rahman, S. K. Sarma, P. A. Boruah and K. Talukdar, "Deep learning based bodo parts of speech tagger", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10308365.

[35] H. Mehta, S. Kumar Bharti and N. Doshi, "Comparative analysis of part of speech (POS) tagger for Gujarati language using deep learning and pre-trained LLM", 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2024, pp. 1-3, doi: 10.1109/INOCON60754.2024.10511678.

[36] K. Talukdar and S. K. Sarma, "Parts of speech taggers for Indo Aryan languages: A critical review of approaches and performances", 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127336.

[37] N. M. Ali, G. H. Ngo and A. L. H. Lan, "Construction of part of speech tagger for Malay language: A review", 25th International Conference on Natural Language Processing (ICNLP), Guangzhou, China, 2023, pp. 253-257, doi: 10.1109/ICNLP58431.2023.00053.

[38] M. Alfian, U. L. Yuhana and D. Siahaan, "Indonesian part-of-speech tagger: A comparative study", 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), Lombok, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICAICTA59291.2023.10390353.

[39] A. Al-Sabri, A. Adam, and F. Rosdi, "Automatic detection of Shadda in modern standard Arabic continuous speech", International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 4-2, pp. 1810-1819, 2018.