University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 9, Issue 2, 2025
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Parts of Speech Tagging for Handwritten Sindhi Sentence using Deep Learning Models

Marya Soomro[1,*], Muhammad Ahsan Raza Mughal[1], Muhammad Khalid Sheikh[2], Azhar Ali Shah[1]

[1]Department of Information Technology & Computer Science Sindh University Campus, Naushahro Feroze ,
[2]Department of Computer Science, National University of Computer and Emerging Sciences (FAST-NUCES) Karachi,

aamarya72@gmail.com, ahsankhan5599z@gmail.com, mkhalidshaikh17@gmail.com, azhar.shah@usindh.edu.pk

Abstract: Part-of-Speech (POS) tagging is an essential task in Natural Language Processing (NLP) that helps identify the role of each word in a sentence. For handwritten low-resource languages, this task becomes more difficult because of the lack of available datasets and differences in individual writing styles. In this study, a deep learning-based approach is developed for POS tagging of handwritten Sindhi sentences. For this purpose, a dataset of handwritten Sindhi sentences was collected and manually labeled with corresponding POS categories. The prepared dataset was then used for training and evaluating the proposed models. The study applies and compares two deep learning models, Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), for automatic POS tagging. The models were evaluated using accuracy, precision, recall, and F1-score metrics. Experimental results demonstrate that BERT achieved better performance compared to LSTM due to its ability to capture contextual information from sentence structures. BERT achieved an accuracy of 87.3% while LSTM achieved 85.1%. The proposed work contributes to Sindhi language processing by providing a handwritten dataset and a baseline deep learning approach for POS tagging of a low-resource language.

Keywords: Computing Methodologies, Artificial Intelligence, Sindhi POS Tagging, Deep Learning Models;

I. INTRODUCTION

Sindhi language is the official language of the Sindh province of Pakistan and also recognized as a scheduled language in India. Spoken by more than 40 million people in Sindh and other parts of the globe, Sindhi language inherits its roots from the Indus Valley Civilization and yet undeciphered Indus Script. From Indus Script to present day Perso-Arabic and Devanagri, the writing systems for Sindhi language have changed over the time including Hatwanki, Gurmkhi, Khudawadi, Shikarpuri, Lohanaki, Khojki and others. This study is based on the Perso-Arabic script (Naskh) which consists of 52 letters which are combination of basic Persian and Arabic letters as well as it uses diacritical marks and dots to modify the basic letters to represent the specific sounds of the Sindhi language. Besides having 46 consonant phonemes and 10 vowels, Sindhi language has a variety of dialects including Vicholi (central dialect spoken in Hyderabad and other central districts), Utradi (northern dialect spoken in Larkano, Shikarpur, Sukkur and Kandiaro), Lari (southern dialect spoken in Karachi, Thatta, Sujawal, Tando Muhammad Khan and Badin), Siroli (siraiki dialect spoken in Jacobabad and Kashmore), Lasi/Firaqi (Balochistani dialect spoken in Lasbela, Hub and Gawadar districts), Thareli (Thari dialect spoken in in the western part of Jaisalmer district of Rajasthan), Bhili (spoken by the aboriginal Sindhi communities including Bhil and Meghwar). At present, most of the Sindhi literature including the mystic poetry of the great Sindhi poet Shah Abdul Latif Bhitai, as well as Sindhi text books are written in Vicholi (centeral) dialect with Perso-Arabic script. Sindhi Parts-of-Speech (PoS) include noun, pronoun, adjective, verb, adverb, preposition, conjunction and interjection. In addition to masculine, feminine, singular and plural, Sindhi noun distinguishes five cases (role as subject, object, etc.) including ablative, locative, nominative, oblique and vocative. The declension including the variety in the form of the Sindhi language PoS are mostly identified based on the grammatical gender and the final vowel. The vowel can be short vowels including ‘a, i and u’ or long vowels including ‘aa, ee, o and oo’ or their nasal sounds. Mostly the -a stems are identified as feminine whereas -o stems are identified as masculine. Examples of various forms of Sindhi PoS including noun, pronoun, verb, adverb, postposition and conjunction are exhibited in Tables 1-4.

Table 1: Various forms of Sindhi noun based on short and long vowel suffices written in Sindhi Perso-Arabic script, Sindhi Roman script and their corresponding meaning in English.

Noun ending with short vowel 'a' mostly feminine	Noun ending with short vowel 'i' mostly Feminine	Noun ending with short vowel 'u' mostly masculine	Noun ending with long vowel 'aa'	Noun ending with long vowel 'ee' mostly feminine	Noun ending with long vowel 'o' mostly masculine
ميز meza table	ڇت chhit~i roof	قلم kalamu pen	هوا havaa air	گهوڙي ghor^ee mare	گهوڙو ghor^o horse
عورت orat~a woman	ريت reet~i tradition	ڪاغذ kaaghhazu paper	فضا Fizaa atmosphere	ڪرسي kursee chair	لفافو lifaafo envelop
نڪ nath~a nose ring	ٻت bhit~i wall	باغ baagu garden	مالها maalhaa necklace	ٻلي b^ilee she cat	ڪيلو kelo banana
زمين Zameena earth	اکي akhi eye	فرش farshu floor	دغا daghhaa fraud	گهڙي ghar^ee watch	دروازو darvaazo door
تصوير t~asveera picture	شڪل shikli face	خط khhat~u letter	سزا sazaa punishment	ڇٽي chhatee umbrella	پنکو Pankho Fan

Table 2: Various forms of Sindhi noun and conjunction written in Sindhi Perso-Arabic script, Sindhi Roman script and their corresponding meaning in English.

Noun singular ending with short vowel 'u'	Noun plural ending with short vowel 'u' changes to 'a'	Noun singular ending with long vowel 'o'	Noun plural ending with short long vowel 'o' changes to 'aa'	Conjunction
مٽ matu claypot	مٽ mata claypots	ٻلو b^ilo tom	ٻلا b^ilaa toms	۽ ain and
اُٺ uthu camel	اُٺ utha camels	گهوڙو ghor^o horse	گهوڙا ghor^aa horses	پر para but
ڳل gulu flower	ڳل gula flowers	صوفو sofo sofa	صوفا Sofa Sofas	يا yaa or
هٿ hath~u hand	هٿ hath~a hands	دروازو d~arvaazo door	دروازا d~arvaazaa doors	جيترليڪ jet~on~ek, either
نڪ naku nose	نڪ naka noses	ڪتو kut~o dog	ڪتا kut~aa dogs	البت albat~a though

Table 3: Various forms of Sindhi pronoun based on first, second and third person written in Sindhi Perso-Arabic Script, Sindhi Roman script and their corresponding meaning in English.

Pronoun first person singular	Pronoun first person plural	Pronoun second person singular	Pronoun second person plural	Pronoun third person singular masculine	Pronoun third person singular feminine
اسين / اسان asee`n/asaa`n we	ٿون too`n you	ٿون too`n you	ٿوهين / ٿوهان t~avhaa`n/ t~avhee`n you	هي / هو he / ho he (person), this/ that (for a thing)	هيءَ / هُوءَ heea / hooa she (person), this /that (for a thing)

Table 4: Various forms of Sindhi adjective, verb and adverb based on singular and plural written in Sindhi Perso-Arabic Script, Sindhi Roman script and their corresponding meaning in English.

Adjective masculine singular	Adjective masculine plural	Adjective feminine singular	Adjective feminine plural	Verb singular	Verb plural	Adverb
ڪاڙو Kaaro black	ڪاڙا Kaaraa black	ڪاري Kaaree black	ڪاريون Kaaryoo`n black	لکڻ likhu write	لکڻُ likhan~u to write	تحڪڙو t~kir^o quickly
نيرو Neero blue	نيرا neeraa blue	نيري neeree blue	نيريون neeryoo`n blue	پڙهه par^hu read	پڙهڻُ parhan~u to read	آهستي aahist~e slowly
اڇو achho white	اڇا achhaa سفيد	اڇي achhee white	اڇيون achhyoo`n white	گهمڻ Ghumu Wander	گهمڻُ ghuman~u to wander	اڪثر Aksar often
سٺو sutho good	سٺا Suthaa Good	سٺي suthee good	سٺيون suthiyoo`n Good	هلڻ Halu walk	هلڻُ halan~u to walk	تنهنڪري Ta`nh`nkare, Hence

In terms of the corpus linguistics, based on the definition and context of a word, the process of marking up the word in a given text corpus as corresponding to a specific PoS is known as PoS tagging or grammatical tagging. The process of PoS tagging is an important area of research in the field of the computational linguistics which uses variety of rule-based and stochastic algorithms for this task each with its own pros and cons. Popular PoS tagging algorithms such as E. Brill's tagger used for the English language also need to be customized for other natural languages including the Sindhi language. The process of PoS is very complex because some words may represent multiple PoS as per different contexts. Even in English language a large number of word-forms are considered to be ambiguous.

Natural Language Processing (NLP) being a subfield of Artificial Intelligence (AI) as well as linguistics focuses on enabling computers to comprehend, interpret, analyze, and generate human language, which opens up a world of possibilities for extensive range of applications, such as machine translation, document analysis, sentiment analysis, chatbots, virtual assistants, text summarization, speech recognition, FAQ systems, text classification and Optical Character Recognition (OCR). Among these applications, document recognition is important and plays a vital role, particularly when dealing with handwritten writings, which present unique challenges than typed documents. Document analysis has been applied to widely spoken languages such as English, Spanish, and Chinese to automatically organize and retrieve information from vast amounts of text. In recent years, there has been a growing interest in applying document analysis techniques to lesser-studied languages, including Sindhi language. Efforts in Sindhi document analysis focus on challenges unique to the language, such as script recognition, morphological richness, and the lack of large annotated datasets. Researchers are developing tools to process automatically Sindhi texts, aiming to enhance information access, digital preservation, and content creation in this language. These efforts include projects that involve OCR for handwritten Sindhi documents and PoS tagging, which helps in understanding the grammatical structure of Sindhi sentences by identifying the function of each word. PoS tagging involves assigning each word in a sentence to the appropriate grammatical category, such as noun, pronoun, adjective, verb, adverb and conjunction. As showcased above, Sindhi language has complex morphology and script variations in handwritten forms, therefore, this process is specially challenging.

Although several studies have focused on Sindhi POS tagging using text-based datasets, limited research exists on POS tagging of handwritten Sindhi sentences. Existing Sindhi NLP resources mainly concentrate on digital text corpora, while handwritten documents introduce additional challenges such as writing variations, character shapes, and noise. Therefore, this research aims to develop a handwritten Sindhi POS tagging framework using deep learning models and evaluate their effectiveness on a custom annotated dataset.

II. Related Work

PoS tagging for Sindhi language has remained an important field of research for quite some time. Various algorithms, models and approaches have been reported and an updated succinct summary of these studies is presented here [1-4]. Adnan Ali Memon and others propose a PoS tagger employing Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, with LSTM providing more successful results. This work is the first to use these deep learning techniques for Sindhi PoS tagging. Utilizing fast text, researchers trained 79,959 Sindhi word vectors, drawn from a corpus compiled from numerous sources including Sindhi literature, stories and poetry. This study is based on 1,459 sentences and 10,584 distinct words which are spilt into 80% for training and 20% for validation. According to results, the LSTM model with an accuracy of 85.80% performed better than GRU model which had an accuracy of 80.77%. Figure 1 exhibits the overall process of this study.

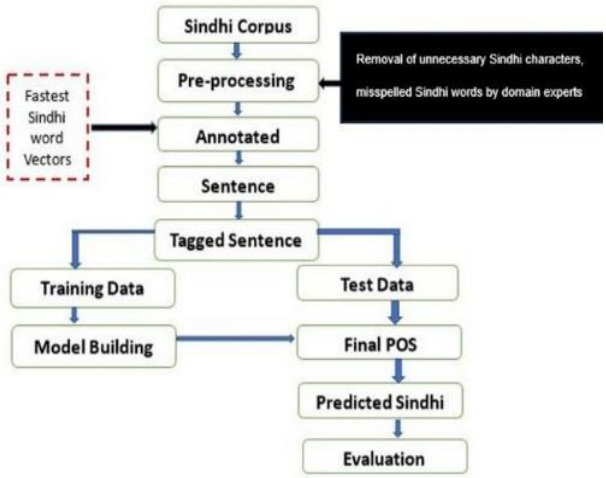

Figure 1: The Architecture of Sindhi POS Tagger (adopted from[1])

Farooq Zaman and others in their article "Leveraging Bidirectional LSTM with CRFs for Pashto Tagging", present LSTM based approach for Pashto Language PoS tagging while focusing on resolving uncertainty [5]. In this study Pashto language corpus of sentences containing concepts of different meanings, were presented with new sentence representation architecture. As per reported results, LSTM model with accuracy of 87.60% for all words excluding punctuation and 95.45% accuracy for ambiguous words performed well as compared to Hidden Markov Model (HMM), which has accuracies of 78.37%, and 44.72% respectively. This result shows how well the LSTM techniques perform in Pashto PoS tagging. Ali Nawaz and others in their article "TPTS: Text Pre-processing Techniques for Sindhi Language" showcase the process of extracting the meaningful information from huge amount of datasets of Sindhi language text using preprocessing [7]. This research is the 1st to use preprocessing model for the Sindhi language, which is designed to do tasks like tokenization, normalization, stop-word removal, stemming and PoS tagging. In this study Sindhi Text Corpus (STC) makes use of 1.5 thousand documents from online sources, for recognizing the linguistic features. TF-IDF approach is employed in this model resulting in 89% accuracy. For text summarization, sentiment analysis and information retrieval in the Sindhi language TPTS is flexible tool for many NLP applications.

Naveen Talpur and others in their article entitled "Researching on analysis and creating corpus from primary level Sindhi language book", highlight the importance of Sindhi language communication and digital interaction, especially on social media and mobile devices. The aim of their study is to analyze primary-level Sindhi text from school books and develop a machine-readable corpus for sentiment analysis. A quantitative approach was employed, with annotations using UPOS, morphological tokens, emotive analysis, and Document Term Matrix. The work of this research starts from understanding the language problem and ends with the development of a comprehensive text corpus for NLP research. The research paper "SiPOS: A Benchmark Dataset for Sindhi Part-of-Speech Tagging" by Wazir Ali and Zenglin Xu, by using the Doccano tool manually annotates over 293K tokens with 16 universal POS categories, and introduces a novel benchmark SiPOS dataset for Sindhi POS tagging [6]. It assesses the dataset using CRF, BiLSTM, and self-attention models, using word and character-level representations for increased POS tagging accuracy. By employing CRF, BiLSTM and self-attention models, it gains a high accuracy of 96.25representations, which makes it precious resource for the low-resource Sindhi language. This research paper "BNLP: Natural language processing toolkit for Bengali" by Sagor Sarker, consists of part of speech (POS) tagging, (NER) facilities and tokenization in a BNLP which is an open-source language processing toolkit for Bengali [8]. Overall Bengali Research communities use BNLP, it has 25k downloads, 138 stars and 31 forks.

The research paper "PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES" by Jamilu Awwalu at el proposed that there are various ways for NLP part of speech (POS) tagging [9]. This research covers both unilingual and multilingual POS tagging which are essential for comprehending sentence structure in many languages. This paper discusses deep learning models like Recurrent Neural Networks (RNN) and long Short –Term Memory (LSTM) networks that have been applied to improve accuracy in multilingual tagging. This review examines the performance, challenges in multilingual contexts, and efficiency in improving language understanding of several POS tagging strategies. The aim in this research paper "Sindhi Language Processing on Online Sindhi NLP Tool" by Erum Naz Sodhar et al, is to create Sindhi NLP tool to perform tasks on Sindhi language. It is consisted on seven

tasks such as: (Lemma (L), Stem Suffix (SS), Stem Affix (SA), Stem (S), Sindhi Parts of-Speech (SPOS), Universal Parts-of-Speech (UPOS) and words (W) from input sentences of Sindhi language [3]. Number of sentences used are 10 and the number of words used are 195 for research to find out root words from Sindhi data. Machine learning algorithms and tools are applied to perform machine learning tasks and get results. According to this research, still no trained data set are available for Sindhi language online. In conclusion, this research focuses on processing the Sindhi language using a specific tool to perform seven tasks: Lemma, Stem Suffix, Stem Affix, Stem, SPOS, UPOS, and words. This study contains 195 words with 10 dataset of Sindhi language to recognize root words. In future for NLP tasks the idea is to use a larger Sindhi dataset and apply unique techniques.

The research paper "Deep learning-based isolated handwritten Sindhi character recognition" by Asghar Ali Chandio el al, labels the problem of handwritten Sindhi text recognition which recommends using a convolutional neural network (CNN)-based summation fusion method and a deep residual network (ResNet) with short cut connections [10]. This model is specially designed for automatic feature extraction and classification of handwritten Sindhi characters. This model is trained using a custom dataset of handwritten Sindhi characters, and while training this model it performed far better than earlier techniques. Making it new to the field, this is the first research to use a deep residual network with summation fusion for handwritten Sindhi text recognition.

The research paper "Sindhi Handwritten-Digits Recognition using Machine Learning Techniques" by Irfan Ali et al, for inquiring the problem of Sindhi Handwritten Digits Recognition this research paper focuses on using a variety of machine learning algorithms [11]. K Nearest Neighbor (KNN), Decision Tree (DT), Multilayer Perceptron (MLP), and Random Forest (RF) Classifier are the algorithms that were used in this research. While identifying the handwritten digits in Sindhi Random Forest Classifier and Decision Tree Models had given better accuracy as compared to other algorithms. For recognizing the digits in Sindhi and evaluating the effectiveness of several ML techniques are the main goal of this study. The research paper "Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches" by Wahab khan et al, is for Urdu POS (Part of Speech) tagging. The main algorithms that are implemented in this research are Conditional Random Field (CRF), Support Vector Machine (SVM), and Hidden Markov Model (HMM) (specifically the bigram HMM), and two variants of Deep Recurrent Neural Networks (DRNN): forward LSTM-RNN and LSTM-RNN with CRF output. [12]. Kalhoro et al. (2025) applied a deep learning-based EfficientNet-B4 model for sugarcane disease detection and showed that deep learning techniques can effectively extract features from complex image data. [13]. Abro et al. (2021) analyzed different machine learning classifiers and highlighted the importance of suitable models for accurate classification tasks. These studies demonstrate the effectiveness of AI-based approaches for automated analysis and classification problems.[14]

Previous studies demonstrate that deep learning models have improved POS tagging performance for low-resource languages. However, most existing approaches focus on typed text datasets rather than handwritten text. Unlike previous work, this study focuses on handwritten Sindhi sentences and investigates the performance of deep learning models under handwriting variability conditions.

III. METHODOLOGY

This section provides the flow diagram of implementing Part of Speech (POS) tagging to improve the understanding and processing of handwritten Sindhi sentences. This research study combines both image processing techniques and deep learning algorithms techniques to create a strong framework for linguistic feature analysis within handwritten data. This research method flow starts with the acquisition of an input dataset for handwritten Sindhi sentences applied POS tagging to a dataset of handwritten Sindhi documents. The dataset contains more than 1000 pictures of annotated Sindhi text, and more than 20 classes.

Table 5: Summary of Handwritten Sindhi Dataset

Property	**Description**
Total images	1000+
Total POS categories	20
Language	Sindhi
Script	Perso-Arabic
Annotation Tool	Roboflow
Training set	57%
Validation set	24%
Testing set	19%

To represent the real-world variances, the dataset contains a variety of handwriting styles. POS categories for nouns, pronouns, adjectives, adverbs, conjunctions, and other grammatical classes were assigned to each sentence. The data covers a wide range of writing styles ensuring comprehensive representation. Data acquisition is the first step of this research approach, the next step of this proposed methodology is pre-processing. The primary functionality of preprocessing step is to improve the image data to enhance image features for further processing or markings that are not relevant and can interfere with text recognition during

the OCR process, the Normalization is one of the preprocessing techniques that are used for adjusting the scale of input data to ensure uniformity and improve model performance in POS tagging tasks, and feature extraction is used for identifying key pattern or characteristics from the text or image data to enhance the accuracy of POS tagging predictions. In addition, handwritten Sindhi sentence images were converted into machine-readable text using Tesseract OCR. The extracted text was manually verified and corrected to ensure annotation quality and reduce transcription errors before POS tagging. The next step is data annotation, it is done by using some python libraries such as labeling. This research has applied manual annotation over the custom created data set, the handwritten sentence images were annotated manually using the Roboflow annotation platform. Each word was given its matching POS category. A Sindhi-specific UPOS tag set was used in the annotation process.

The annotated corpus consists of noun, pronoun, verb, adjective, adverb, and preposition, conjunction, and interjection categories along with their subcategories. The frequency distribution of each tag is presented in Table 6.

Table 6: Distribution of POS Tags in the Annotated Handwritten Sindhi Corpus

UPOS Tag	Tag Type Description	Count
N	Noun	340
N-CM	Common Noun	411
N-PR	Proper Noun	218
N-AB	Abstract Noun	204
N-C	Countable Noun	210
N-NC	Non Countable Noun	190
PR	Pronoun	328
PR-OB	Objective Pronoun	224
PR-IN	Interrogative Pronoun	190
V	V	424
ADV	ADV	204
ADV-TM	ADV-TM	201
ADJ	ADJ	241
ADJ-POS	ADJ-POS	201
PREP	PREP	367
CONJ	CONJ	201
INT	INT	201
V-PRS	V-PRS	403
V-PST	V-PST	223
V-PRS-CONT	V-PRS-CONT	205

The annotated dataset was then converted into a structured format for model training and evaluation. The following stage of this research paper is known as data preparation. It involves some tasks like cleaning, normalizing, and splitting the data into training, validation, and test sets. The model will learn efficiently and be able to generalize successfully to new data if the data is prepared properly. In this step, dataset is divided into three parts: Training, Validation and Testing sets. Training set is 57%, validation set is 24% and testing set is 19%. This session ensures that the model is reliable and works effectively with raw data. Validation set is used to adjust hyper parameters and asses the model's performance during training ensuring that it generalizes properly.

The test set is used to measure accuracy, precision, and other metrics by evaluating the trained model's final performance on unseen data. Next step is Model training, in this step model is trained using some algorithms like LSTM, BERT are selected based on the problem requirements. The multilingual BERT (mBERT) model was chosen because it supports multiple languages including scripts similar to Sindhi and provides contextual word representations. The POS labelled data is used to train these models. To ensure reproducibility of the experiments, the training parameters used for LSTM and BERT models are summarized in Table 7. These parameters were selected to optimize model performance and maintain consistent evaluation.

Table 7: Hyper parameters Used for Model Training.

Parameters	LSTM	BERT
Epochs	50	5
Batch Size	32	16
Learning Rate	0.001	2e-5
Optimizer	Adam	AdamW
Loss function	Cross Entropy	Cross Entropy

Finally, the model evaluation, its performance is tested on the unseen test dataset. For calculating the performance of a model to check how well it predicts POS tags, metrics like accuracy, precision, recall and F1-score are used. Figure 3 shows that how each step of our methodology is important for this research giving robust and trustworthy results that contribute to the further advancement in the domain.

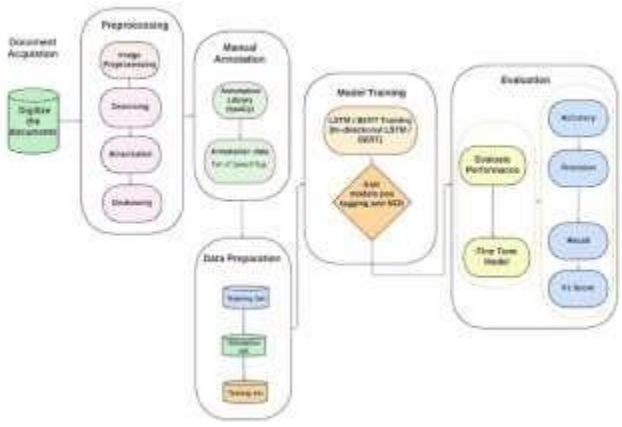

Figure 3: Proposed Research Methodology

IV. RESULTS

This section presents the results obtained from the experiments designed to investigate the POS tagger system accuracy, Precision, Recall and F1 Score using two deep learning techniques the first one is LSTM and second one is BERT model, this research experiments on the created Custom dataset it demonstrates that in order to achieve high accuracy without over fitting, current experiment of our research results is discussed below. In assessing the primary performance of our POS tagging models for handwritten Sindhi sentences, the BERT and LSTM models determine different strengths and restrictions across key metrics, specifically accuracy, precision, recall, and F1-score.

Table 8: Performance Comparison of POS Tagging Models on Handwritten Sindhi Sentences

Algorithms	Accuracy	Precision	Recall	F1-Score
LSTM	85.1%	83.1%	82.3%	82.7%
BERT	87.3%	88.5%	88.7%	88.1%

The overall results emphasize the value of combining BERT's context-rich embedding's with a tagging model for handwritten, low-resource languages like Sindhi, offering understandings into future improvements that could further enhance tagging accuracy. In figure 4 and 5 the output result of the model have been visualized using matplotlib library for better understanding of the results.

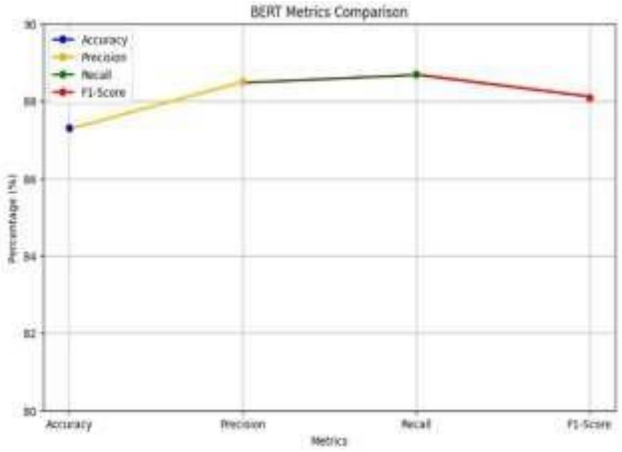

Figure 4: BERT Metrics Comparison

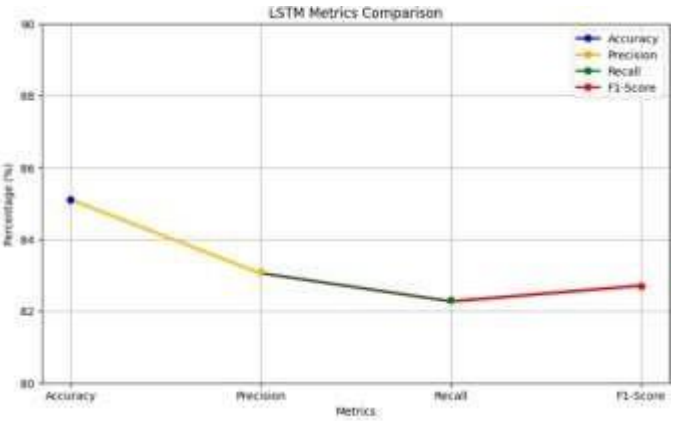

Figure 5: LSTM Metrics Comparison

V. DISCUSSION

BERT achieved higher performance because it uses contextual embedding's that consider the relationship between words in a sentence. Unlike LSTM, which processes sequential information step-by-step, BERT captures bidirectional context, making it more effective for handling complex grammatical structures in handwritten Sindhi sentences. The obtained results are comparable with previous Sindhi POS tagging studies. Earlier research mainly focused on digital Sindhi text, while this study addresses handwritten Sindhi sentences, which introduces additional challenges due to handwriting variation.

VI. CONCLUSIONS

In conclusion, this research study demonstrates that the use of deep learning techniques such as BERT and LSTM result in effective results for the POS tagging system in handwritten Sindhi sentences which is a low-resource language. The BERT model resulted in higher accuracy and F1-scores, primarily outperforming in context-rich tagging tasks whereas the LSTM model performed adequately for simpler sequential tagging. In totality, the BERT and LSTM models succeeded in solving unique challenges of handwriting variability in Sindhi suggesting deep contextual embedding's meaningfully enhances model performance parameter for compound linguistic tasks. These findings show that advanced NLP methods can be useful for developing POS tagging systems for under-resourced languages. The proposed approach can be further improved and applied to other handwritten language datasets in future research. Future studies can further improve these models by training them on larger and more diverse handwritten datasets to make them more reliable for different handwriting variations.

VII. Acknowledgment

The authors acknowledge the support provided by the University of Sindh, Jamshoro, during this research work. The authors also appreciate the guidance and valuable

feedback received from their supervisors and reviewers, which helped improve the quality of this study.

VIII. REFERENCES

[1] Adnan A. Memon, Saman Hina, Abdul K. Kazi, Saad Ahmed. 2024. Parts-of-speech tagger for Sindhi language using deep neural network architecture. MUET Research Journal 43, 3 (July 2024), 47-55. https://doi.org/10.22581/muet1982.2768

[2] Mutee R. Arain. 2015. Towards Sindhi Corpus Construction. Linguistics and Literature Review 1, 1 (March 2015), 39-48. https://doi.org/10.32350/llr/11/04

[3] Erum N. Sodhar, Hina Bhanbro, Zira H. Amur, Akktar H. Jalbani, Abdul H. Buller. 2020. Sindhi Language Processing on Online SindhiNLP Tool. University of Information and Communication Technology 4, 3 (October 2020), 139-142. https://sujo.usindh.edu.pk/index.php/USJICT/article/view/2879

[4] Vijay Rowtula and Praveen Krishnan. 2018. POS Tagging and Named Entity Recognition on Handwritten Documents. In Proceedings of the 15th International Conference on Natural Language Processing. NLP Association of India. 82-86. https://aclanthology.org/2018.icon-1. 11/

[5] Farooq Zaman, Onaiza Maqbool, and Jaweria Kanwal. 2024. Leveraging Bidirectional LSTM with CRFs for Pashto Tagging. ACM Trans. Asian Low Resour. Lang. Inf. Process. 23, 4, Article 58 (April 2024), 17 pages. https://doi.org/10.1145/3649456

[6] Wazir Ali, Zenglin Xu, and Jay Kumar. 2021. SiPOS: A Benchmark Dataset for Sindhi Partof-Speech Tagging. In Proceedings of the Student Research Workshop Associated with RANLP 2021, pages 22–30, Online. https://aclanthology.org/2021.ranlp-srw.4/

[7] Nawaz, Ali Nawaz, Muhammad Shaikh, Noor Rajper, Samina Baber, Junaid Khalid, Muhammad. (2023). TPTS: Text Pre-processing Techniques for Sindhi Language. Pakistan Journal of Emerging Science and Technologies (PJEST). 4. 1-12. 10.58619/pjest.v4i3.89

[8] Sagor Sarker (2021), BNLP: Natural language processing toolkit for Bengali language, CoRR, abs/2102.00405. https://arxiv.org/abs/2102.00405

[9] Awwalu, Jamilu Abdullahi, Saleh Evwiekpaefe, Abraham. (2020). PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES. FUDMA Journal of Sciences. 4. 712-721. 10.33003/ fjs-2020-0402-325

[10] Ali, Asghar. (2020). Deep learning-based isolated handwritten Sindhi character recognition. Indian Journal of Science and Technology. 13. 2565-2574. 10.17485/IJST/v13i25.914

[11] Ali, Irfan Ali, Insaf Guriro, Subhash Khan, Asif Raza, Syed Qureshi, Basit Bhatti, Priha. (2019). Sindhi Handwritten-Digits Recognition Using Machine Learning Techniques. International Journal of Computer Network and Information Security. VOL.19. pp. 195-201.

[12] Khan, Wahab Daud, Ali Khan, Khairullah Nasir, Jamal Basheri, Mohammed Aljohani, Naif Alotaibi, Fahd. (2019). Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2897327

[13] A. A. Kalhoro, R. Husain, and H. Shaikh, “Deep Learning for Sugarcane Disease Detection: A Field-Validated EfficientNet-B4 Approach,” *University of Sindh Journal of Information and Communication Technology*, vol. 9, no. 1, pp. 28–38, 2025.

[14] A. A. Abro, A. A. Khan, M. S. H. Talpur, I. Kayijuka, and E. Yaşar, “Machine Learning Classifiers: A Brief Primer,” *University of Sindh Journal of Information and Communication Technology*, vol. 5, no. 2, pp. 63–68, 2021.