IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 April 2024, accepted 29 May 2024, date of publication 11 June 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412175

RESEARCH ARTICLE

A Deep Learning-Based Approach for Part of Speech (PoS) Tagging in the Pashto Language

SHAHEEN ULLAH[1], RIAZ AHMAD[1], ABDALLAH NAMOUN[2], (Member, IEEE), SIRAJ MUHAMMAD[1], KHALIL ULLAH[3], IBRAR HUSSAIN[4], AND ISA ALI IBRAHIM[5]

[1]Department of Computer Science, Shaheed Benazir Bhutto University (SBBU), Sheringal, Upper Dir, Khyber Pakhtunkhawa 18050, Pakistan
[2]Faculty of Computer Science and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia
[3]Department of Software Engineering, University of Malakand (UOM), Khyber Pakhtunkhawa, Pakistan
[4]Department of Computer Science and Information Technology, University of Malakand (UOM), Dir Lower, Chakdara 18800, Khyber, Pakhtunkhawa, Pakistan
[5]Department of Cybersecurity, School of Information and Communication Technology, Federal University of Technology Owerri, Owerri 460114, Nigeria

Corresponding author: Ibrar Hussain (ibrar@sbbu.edu.pk)

ABSTRACT A fundamental task in natural language processing (NLP) is part of speech (PoS) tagging. PoS tagging is crucial to many NLP applications, including question answering, machine translation, syntactic parsing, speech recognition, and semantic parsing. PoS tagging is a task for labeling sequences in which a tagger/ system tags each word with its appropriate part of speech label. In NLP, PoS tagging is often considered as a language-specific task. Similarly, Pashto is a language that has not been explored regarding PoS tagging. Therefore, this research focuses on the PoS tagging considering the Pashto language and provides a baseline accuracy. The research has twofold benefits. First, it introduces a Pashto tag set that contains 2,81,205 words of the Pashto language. All these words are tagged with 17 unique PoS tags. Second, it proposes a deep learning-based model by examining classic Recursive Neural Networks (RNN) and Bidirectional Long Short Term Memory Networks (BLSTM). The results show promising performances when used with the word embedding technique. The proposed approach achieved 98.82% accuracy as a baseline on the test dataset by using the BLSTM model along with word embedding.

INDEX TERMS Artificial intelligence, document image analysis, handwritten text, natural language processing, optical character recognition, speech recognition, standard dataset.

I. INTRODUCTION

A natural language is often referred to as an ordinary language used by humans to speak or write for common communication [1]. Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) [2] that deals with incorporating the ability and comprehension of natural languages into machines. However, some of the tasks in NLP are language-specific and need specific treatment regarding language. One such tasks in NLP is Part of Speech (PoS) tagging. In PoS tagging, machines tag each component of a natural language with its appropriate label. These labels usually represent a certain part of speech, for example; Verb, Noun, Adjective, Pronoun, etc. There are 48 unique PoS tags regarding the English language [3]. However, this number may vary and is language dependent.

Similarly, Pashto is a low-resource language with little work in the field of NLP. Pahsto language is spoken by 50 million people across the world [4]. It is famous for its rich culture and literature associated with poetry and music [5]. However, the Pashto language in the field of NLP needs further investigation, especially for PoS tagging in the Pashto language.

PoS tagging [6] is an important activity of NLP. Parts of speeches are the generic names/ tags used for an individual word in a sentence. Major parts of speech are Noun as NN,

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

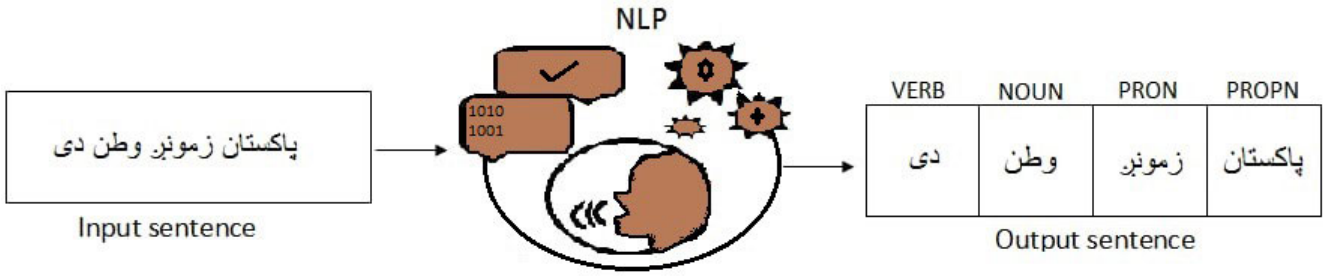


FIGURE 1. PoS tagging of a sentence taken from the Pashto language.

Verbs as VB, adjectives as JJ, pro-noun as PRP etc. In the PoS tagging, we choose a suitable tag for each word in a sentence. Figure 1 represents a sentence of a Pashto language along with their corresponding POS tags. Additionally, PoS tagging is an essential part of NLP applications and guides individuals to identify the use of a word in a sentence. Grammatical categories such as tense, numbers (singular/ plural), nouns, verbs, adjectives, etc. can also be distinguished by using the Part of Speech tagger. Moreover, PoS tagging is very beneficial for named entity recognition (NER), building parse trees, and extracting a relationship among words.

There are several approaches to generating a PoS tagger. Such approaches are either using a Rule-based approach or using Deep learning models. Moreover, Hidden Markov Model (HMM) [7] based approaches are statistical and could give better results. However, these methods are still lacking good accuracy compared to deep learning methods. Recently, Long Short Term Memory (LSTM) based PoS taggers have shown significant performance in achieving state-of-the-art accuracy in various languages [8]. LSTM has feedback connections, a variant of recurrent neural networks (RNNs) [9]. It is capable of learning long-term dependencies, especially in sequence prediction problems. Thus, our proposed model/PoS tagger will be based on LSTM-based neural networks.

This research contributes to two major areas in NLP regarding the Pashto language. Firstly, it provides a Pashto Tag-set containing 281205 words taken from the Pashto Handwritten Text Imagebase (PHTI) [10] that can be used in for variety of Pashto NLP task such as word segmenter, Pashto dictionary and resolving the space anomaly issues. These words are in text form, and we have tagged 5000 unique words by considering the top most used words in the Pashto language. Secondly, this research also provides and suggests a deep learning model integrating RNN [11] and LSTM [12] architectures. Furthermore, the simple RNN and LSTM based models are examined with and without word embedding showing and validating the positive impact of word embedding in NLP tasks especially in PoS tagging.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 discusses the proposed methodology. Section 4 represents the experimental setup. Section 5 includes the results and discussion, and Section 6 presents the conclusion and future directions for the research.

II. RELATED WORK

There has been an enormous work regarding the research in the area of PoS tagging. However, major work focuses languages that are considered as rich resource languages including English, French, Arabic, Chinese etc. However, in this article, we addressed the very close work that can be adapted for the PoS tagging of Pashto language.

A hidden Markov model based PoS tagger was presented by [13] targeting the Persian corpus. In their work, the fundamental elements of Persian morphology are presented and developed. Their approach is used in simulations on both homogeneous and heterogeneous Persian corpora to assess the correctness of the suggested approach. They achieved an accuracy of 98.1% on the Persian corpus.

Hasan et al. [14] compared different Part of speech taggers models like the n-gram base model, Hidden Markov Model (HMM), and Transformation based tagging (TBLs) for South Asian languages. He trained a corpus of 20, 000 words and acquired a performance of 90% accuracy. He also showed that when the size of the corpus is increased then the accuracy will also increase.

Mohammed et al. [15] developed a generic PoS tagger and word embedding for the Somali language by using techniques like HMM and Conditional Random Field (CRF) [16]. They also used Neural Network and achieved a state-of-the-art PoS tagger by obtaining 87.51% accuracy on a 10-fold cross-validation.

Makarenkov et al. [17] presented a novel system by using machine learning to perform Lexical substitutions and grammatical error correction. The researchers applied Long Short-Term Memory (LSTM) [18] as tagger. Their model was able to scan the input in both directions The purpose of the study was to address the challenges identified by the scholars, who have English as a second language.

Marquez et al. [19] presented a machine learning based technique for constructing statistical language models for PoS tagging. They directly applied the learned models to a quick, easy tree-based tagger and got reasonable results. Additionally, they created a customizable relaxation-labeling based tagger by merging the models with n-gram statistics.

They demonstrated how both models effectively cooperate to produce better results. When huge training corpora are compared to reasonably small training sets, it becomes clear that the combination of the learnt tree-based model with the best n-gram model produces the best results in both cases.

Sarmady et al. [20] studies done for Persian text part of speech tagging using Markov Model, Memory based, and Maximum Likelihood techniques. The taggers were trained on 85% of the POS corpus created for these experiments, and they were tested on the remaining 15%. The findings demonstrate that, without any prior linguistic training, we are able to develop a workable statistical Part of Speech tagger for the Persian language.

Baig et al. [21] proposed a novel PoS-tagged dataset created from Urdu tweets in this research paper, along with its tagging system. They carried out an experiment where they assessed how two pre-trained Urdu taggers performed on well-edited Urdu text and Urdu tweets. The performance of these taggers on Urdu tweets was significantly lower, according to the results. Researchers described the creation of a manually tagged dataset of 500 Urdu tweets, the consistency of which was assessed using five-fold cross-validation.

Alrajhi and Elaffendi [22] presented PoS tagging for the Arabic Language using the deep learning-based approach. Additionally, the LSTM method was used during the study. A machine learning technique was used to train the model first and then test it on data. They used the Arabic morphemes taken from the corpus known as the Quranic Arabic Corpus (QAC). Their system attained 99.72% for tagging QAC. Furthermore, 99.18% for labeling words improved the performance compared to the previous studies.

Pashto is considered a low-resource language and there is little work for the Pashto language regarding PoS tagging. In this context, Rabbi et al. [23] developed a Pashto tag-set for the Pashto language. They proved that the tag-set is very suitable for generating PoS tagging for the Pashto language. The researchers showed that it is vital to design a tag-set for producing Part of Speech tagging for any language. Similarly, the researchers applied the Pashto tag-set for making Part of Speech tagging in the Pashto language by using a rule based method.

It can be observed that very limited work has been conducted specifically addressing the PoS tagging in the Pashto language. The major reason seems to be the unavailability of labeled data. Therefore, this work focuses the mentioned gap and contributes in terms of new tag-set and a baseline classifier based on deep leering models.

III. PROPOSED METHODOLOGY

To achieve the objectives of PoS tagging in the Pashto language, we applied a neural network model by using a deep learning [24] approach. First, we created a Pashto tag-set, then annotated [25] the Pashto tag-set. After the annotation, the deep learning models are trained on the corpus and subsequently tested on the test set. The next section describes the creation process of the new tag-set.

A. PASHTO TAG-SET CREATION

The Pashto textual data is taken from the ground truth file associated with the Pashto Handwritten Text Image base (PHTI) [10], which consists of 4, 20, 961 words. The ground-truth file only contains the textual data or labels that annotate handwritten text-line images. Further, we converted this PHTI text file into a separate value (CSV) file using a Python script to find the most frequently occurring Pashto words. The CSV file was then arranged in descending order, where the most frequent word of the Pashto was indexed on the first line. In this way, the top 5000 unique Pashto words were isolated and manually annotated using the help of Pashto dictionaries like Daryaab [1] and Google translator. The Google translator is only used when two similar words were having different meanings. The major annotation was done manually by considering the context of the text.

On the basis of these 5000 unique words, the original file (PHTI.txt) was tagged with suitable PoS tags. As a result, a total of 281205 words of Pashto were tagged, and the new file was named *proposedTAGSET.txt*. It should be noted that the CSV file has only 5000 unique Pashto words, where they are ordered based on frequency. On the other hand, in the final file i.e., *proposedTAGSET.txt*, the words are retained in their natural flow of reading and writing. Due to this coherent nature of our proposed tag-set, it is highly compatible with the exploration of the structure and composition of the Pashto language. Figure 2 explains the procedure after the CSV file creation to the final Target file. Further, the tag-set contains 16 PoS tags in the Pashto language. It should be noted, that there may exist more than 16 PoS tags in the Pashto language. Our findings could not be generalized as we believe that the numbers may be more. However, finding the exact number may need another research and could be a good topic for the researchers working in the field of linguistics. In our study, the PoS tag i.e. "x" can be assumed for other non labelled tags. Table 1 shows the total number of part of speech tagging (PoS). The proposed tag-set can be downloaded using the given url.[2]

B. PROPOSED MODEL

In this research, we have examined and adapted the two most reputable deep learning models [26] like Recurrent Neural Network (RNN) [27] and BLSTM [28] models. Further, these models are also checked [29] with and without word embedding weights. The following major parts provide explanations of these two neural network models.

1) RECURRENT NEURAL NETWORK

The conventional feed-forward neural networks [30] have limitations, as they are unable to use the output back as input. In contrast, the RNNs can use the output of the model back to its input in the next step. In other words, they have recurrent connections which are used to feed back the output as an

[1] https://www.pukhto.net/books/daryaab-dictionary
[2] https://github.com/adqecsbbu/PHTI/blob/main/proposedTAGSET.txt

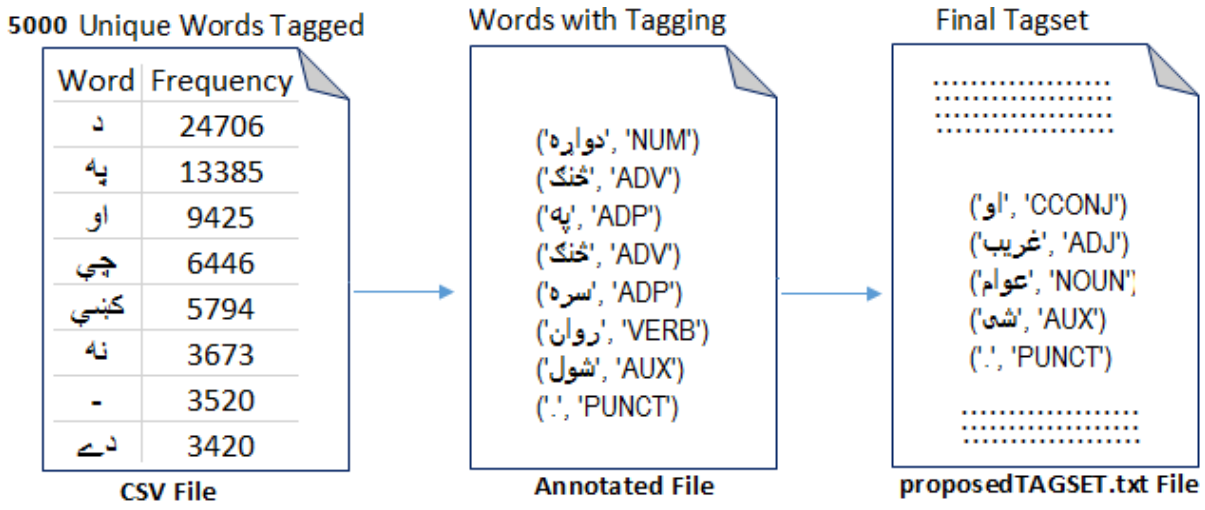


FIGURE 2. The creation process of a final tag-set file after the CSV file that leads to the final creation of Pashto Tag-set.

TABLE 1. Examples of 16 PoS tags regarding Pashto proposed tagset.

S.NO	Category	PoS	Example
1	Noun	NOUN	(ښار, NOUN)
2	Conjunction	CCONJ	(مګر , CCONJ)
3	Adpostion	ADP	(باندی, ADP
4	Verb	VERB	(وویل, VERB)
5	Pronoun	PRON	(هغه , PRON)
6	Adjective	ADJ	(ښکلی, ADJ)
7	Auxiliary	AUX	(وي , AUX)
8	Adverb	ADV	(اوس , ADV)
9	Other	X	(تتار , X)
10	Proper Noun	PROPN	(پیښاور, PROPN)
11	Number	NUM	(۲۲ , NUM)
12	Punctuation	PUNCT	(. , PUNCT)
13	Particle	PART	(اتها , PART)
14	Article	ART	(او , ART)
15	Interjection	INTJ	(واه واه , INTJ)
16	Symbol	SYM	($, SYM)

input. Such recurrent connection helps to learn the past and to predict the future. Therefore, RNNS are famous for sequence learning problems [31]. As PoS tagging is one of the sequence learning problems, we have opted to use the very basic model of RNN [32].

2) BIDIRECTIONAL LONG SHORT TERM MEMORY (BLSTM)

Another approach that is very popular and has been used in many sequence classification problems is based on an advanced variant of RNN known as Bidirectional Long Short Term Memory (BLSTM) networks. As mentioned

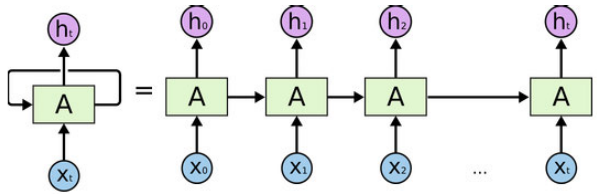


FIGURE 3. Typical RNN and its expansion over time step [33].

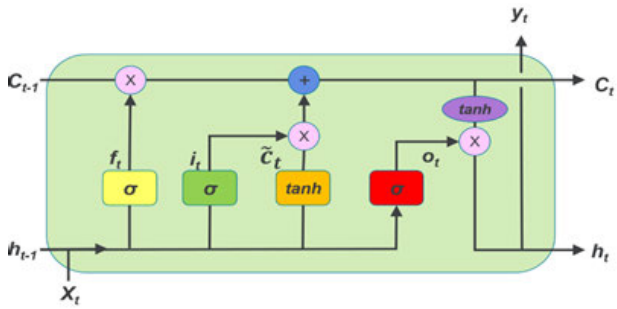


FIGURE 4. Basic diagram of LSTM [36].

in the previous section, RNNs can learn from the past. However, as long as the past goes away, the learning in RNN models fades. Because storing and learning the long-term context is impossible for the classic RNNs, this issue is also known as the gradient vanishing problem [34]. The reason is that they do not have any memory mechanism that can associate long-term dependency with future input. Shumdeber et al. [18] developed the LSTM models to cope with this issue. LSTM-based approaches have gained tremendous reputations and have achieved better performances on various sequence-related benchmarks [35]. Figure 4 shows the architecture of the BLSTM model. In terms of architecture, BLSTM is a classical RNN with extra gates. These gates regulate the learning mechanism of the model and decide when to read, forget, and write the new information in the BLSTM memory. The following text describes the gates and their functionality in a brief manner.

Input gates that are based on previous outputs regulate the information that will be transmitted through memory cells. The coming input is checked by the input gate. The equation is expressed mathematically in equation 4.1.

$$i_t = a\,(w_i[h_{t-1},\, x_t] + b_i) \tag{1}$$

The forget gate is used to forget certain information and keep some crucial information when the context is changing. It retains all information when the context is constant. The forget gate can be explained in math by the given equation 2.

$$f_t = a(wf\,[\,h_{t-1},\, x - t], b_f) \tag{2}$$

The Output gate controls the output at the current time step. It decides whether the LSTM unit will produce output at the current time steps or nothing. Equation 4.3 describes the output gate in a mathematical manner.

$$o_t = o\,(w_o[h_{t-1},\, x_t] + b_o) \tag{3}$$

3) WORD EMBEDDING

Word embedding [37], [38], [39] is a classical representation of words and has been used significantly in many NLP applications. In this representation, each word is represented in a vector form with a certain number of dimensions. The dimensions may be 100, 200, or 300, and their value depends on the depth and vocabulary size of the target language. In our case, the dimension is taken as 300 for word embedding in the Pashto language. Each value in the vector represents how close or away is a particular word with the other word. The higher the value, the more will be that word relevant and vice versa.

A model used in natural language processing is called World2vec [40] model. The Word2Vec uses the dataset from the text input and produces a vector as a result. Machines cannot immediately understand text-based data. To turn text data from a corpus [41] into a numerical form that a machine can readily understand, the word2vec model [42] will be needed. In this research, We have used all the above-mentioned models, especially the simple RNN and RNN with LSTM. Further, we have examined the impact of word embedding along with the BLSTM and RNN-based models.

IV. EXPERIMENT

The models that we have discussed in Section III are examined by conducting the following procedure for experimentation. However, before going into details, it is important to explain the split mechanism of our newly created tagset. Therefore, the following section describes the split of the dataset into train, test, and validation sets.

A. DATA SPLIT

To exploit the supervised learning approach [43] we will need data in training, testing, and validation sets. It is an essential requirement for the conduction of the proposed experimental procedure. For this purpose, the final tagset file i.e. "proposedTAGSET" contains a word with its appropriate PoS tag on each and every different line. The overall file is split into 70% training set, 15% into test set, and the remaining 15% into validation set.

B. EXPERIMENT ON RNN WITHOUT WORD EMBEDDING

In this experiment, we have examined the classic RNN without word embedding technique on our newly created dataset. The experiments are composed of a grid search to find the optimal size for RNN units. Table 2 shows the overall experiments and the final RNN model contains 64 RNN units.

C. EXPERIMENT ON RNN WITH WORD EMBEDDING

The previous model (i.e. RNN with 64 Units) is also examined with the combination of word embedding [44]. For word embedding, we used Fasttext [3] resources for extracting the Pashto word embedding weights. We utilized a command line interface to obtain the Fasttext library and downloaded

[3]https://fasttext.cc/docs/en/crawl-vectors.html

TABLE 2. The grid analysis of RNN model without world embedding weights, with optimizer adam where Val represents validation and Acc represents accuracy.

RNN Units	Batch Size	Epochs	Val Loss	Val Acc	Test Loss	Test Acc
2	16	8	0.61	0.80	0.50	0.80
4	20	15	0.41	0.86	0.35	0.82
8	25	30	0.27	0.87	0.25	0.84
16	30	50	0.21	0.88	0.21	0.86
32	50	80	0.19	0.89	0.21	0.87
64	128	100	0.17	0.91	0.17	0.88

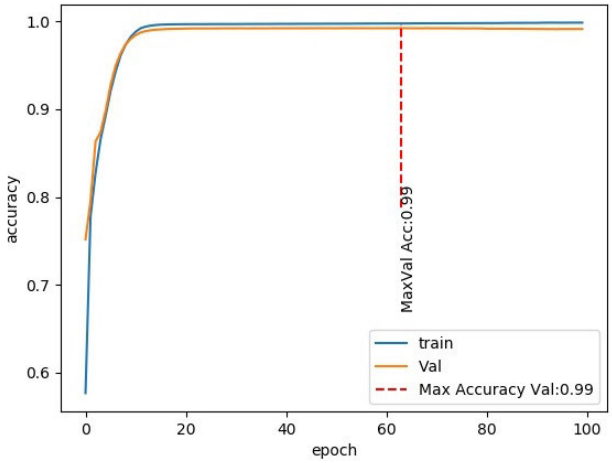


FIGURE 5. Training process of RNN model with word embedding. The max accuracy on test set is 99.0%.

TABLE 3. Grid analysis of RNN model with world embedding weights where Val represents validation, Acc represents accuracy.

RNN Unit	Batch Size	Epochs	Val Loss	Val Acc	Test Loss	Test Acc
2	16	8	0.64	0.83	0.60	0.85
4	20	15	0.43	0.89	0.40	0.89
8	25	30	0.30	0.91	0.28	0.91
16	30	50	0.26	0.92	0.24	0.92
32	50	80	0.24	0.93	0.23	0.93
64	128	100	0.20	0.94	0.19	0.94

a Pashto vocabulary file i.e. *cc.ps.300.vec Size:2.91GB*. The weights were extracted via Tensorflow and were used as trainable weights [45] in the initial layer of the RNN model. Table 3 shows the overall experiments carried out as grid search; we can see that the accuracy is improved by 5.31% while using the word embedding along with the RNN model with optimizer adam. Figure 5 shows the overall process during training for RNN with word embedding.

D. EXPERIMENT ON BLSTM WITHOUT WORD EMBEDDING

As discussed in Section III, BLSTM-based models are highly effective when used for NLP applications. Therefore, in this experiment we have examined the power of BLSTM but without word embedding. Table 4 shows the grid search for

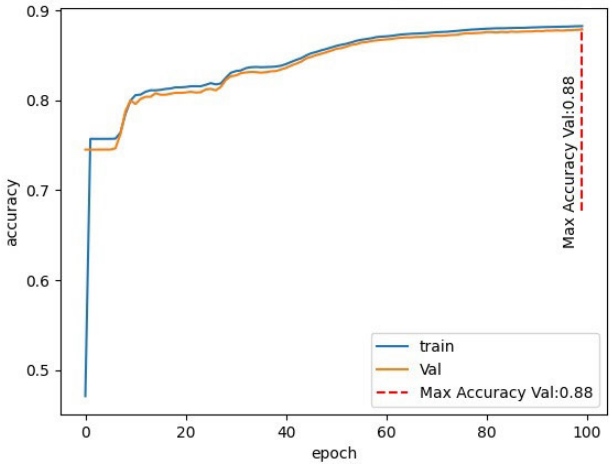


FIGURE 6. BLSTM model used without word embedding. The max accuracy is 88% on test set.

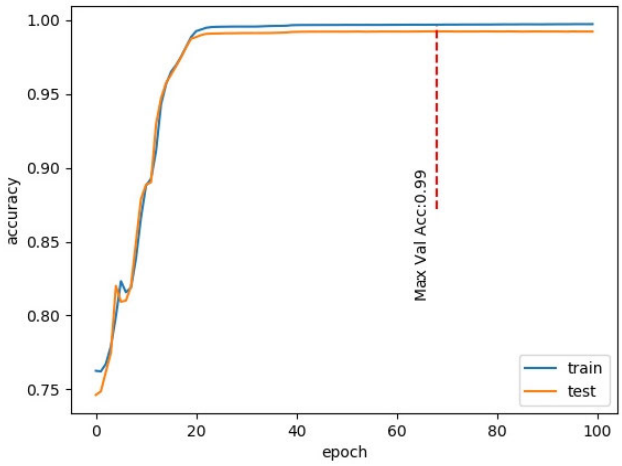


FIGURE 7. BLSTM with weights of word embedding. The max accuracy obtained on test set is 99%.

finding the optimal size for LSTM units for the proposed model with optimizer adam. It is shown that the LSTM-based approach [46] achieved 88.00% accuracy on the test set, which is comparatively less than classic RNN. However, to deepen the LSTM layers up to 128 LSTM units we achieved better performance by achieving 94.27% accuracy on the test set. Figure 6 illustrates the overall training process.

E. EXPERIMENT ON BLSTM WITH WORD EMBEDDING

In this experiment, the BLSTM model with 64 units is combined with word embedding using the same procedure mentioned in Section III. The proposed model outperforms the former models by a significant margin. Results show that BLSTM+Word Embedding has achieved 98.8% accuracy on the test set. Similarly, increasing the number of epochs and LSTM units improves the accuracy of the test set. For example, using 128 LSTM units with word embedding and training the model up to 100 epochs could lead the accuracy

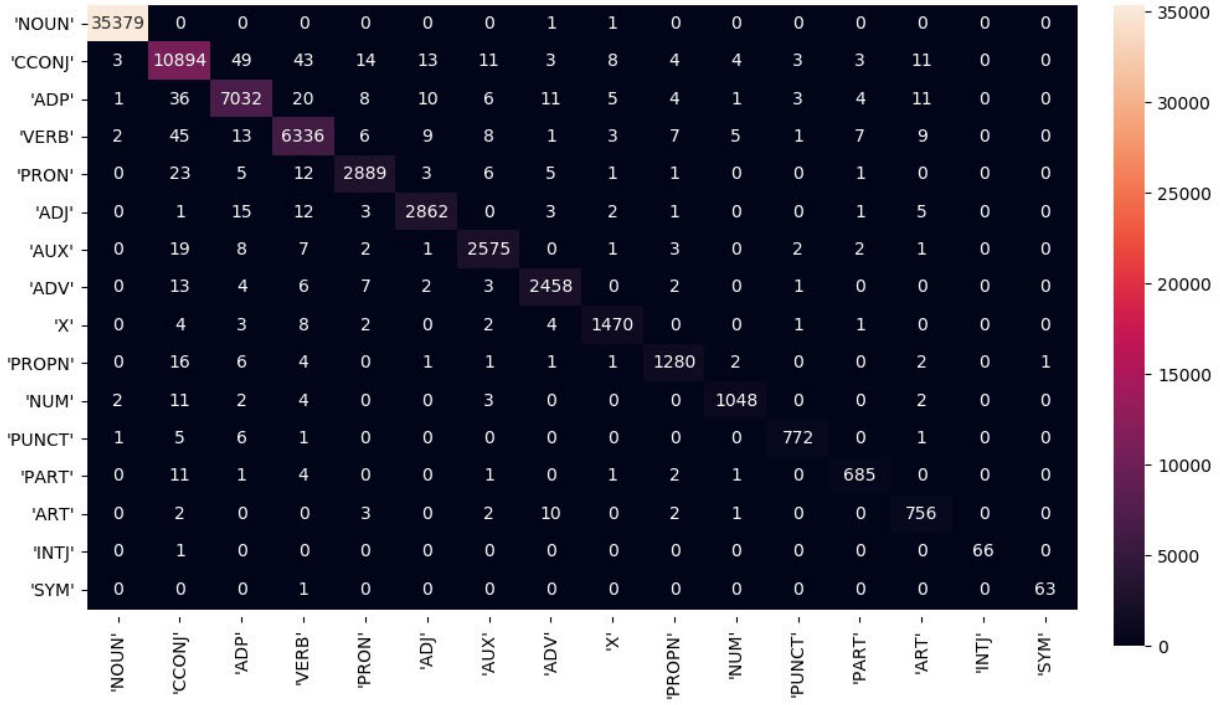


FIGURE 8. Confusion matrix computed on BLSTM with Word embedding.

TABLE 4. Grid analysis of BLSTM model with non-embedding weights where Val represents validation dataset and Acc represents accuracy.

LSTM Unit	Batch Size	Epohcs	Val Loss	Val Acc	Test Loss	Test Acc
2	16	8	0.69	0.85	0.65	0.86
4	20	15	0.47	0.86	0.43	0.87
8	25	30	0.38	0.90	0.35	0.90
16	30	50	0.23	0.93	0.22	0.93
32	50	80	0.19	0.94	0.19	0.94
64	128	100	0.23	0.94	0.22	0.94

TABLE 5. Grid analysis of LSTM model with embedding weights where Val and Acc represents validation and accuracy respectively.

LSTM Unit	Batch Size	Epochs	Val Loss	Val Acc	Test Loss	Test Acc
2	16	8	0.58	0.86	0.55	0.89
4	20	15	0.29	0.92	0.28	0.92
8	25	30	0.78	0.98	0.08	0.98
16	30	50	0.05	0.99	0.06	0.98
32	50	80	0.09	0.99	0.06	0.98
64	128	100	0.05	0.99	0.06	0.98

up to 98.82% on the test set. Table 5 shows the grid search for finding the optimal size for LSTM units for the proposed model with optimizer adam.

Finally, we can compare the results and it is observed that LSTM with word embedding has a better performance compared to RNN with and without word embedding [46].

V. RESULTS AND DISCUSSIONS

As it is the first time to evaluate the newly proposed Pashto Tag set. Therefore, the direct comparison with the other models is not possible. However, we have examined well known RNN models including LSTM. Further, we used word embedding technique [47] which has shown promising results when added to RNN as well as to LSTM model. Table 6 shows the overall comparison of the proposed models. It is clear, that LSTM with word embedding approach has shown improvement. In the next section, we have discuss the major finding using confusion matrix.

A. FINDING AND DISCUSSION

We computed the confusion matrix on the test set of our Pashto tagset. The confusions were estimated by using BLSTM with word embedding model [48]. Figure 8 shows the overall confusion related to PoS tags in the test set. Keenly observing each of confusions, we can conclude that in the Pashto language Noun is abundant and the PoS tagger has learned a lot while classifying the Noun as ''noun''. The overall performance is satisfactory, as the test set is unseen for these models.

According to a review presented in [49], reports that regarding PoS tagging the deep learning based approaches are dominantly used and are producing better performance compared to other approaches including Naive Bayes, Hidden Markova Model (HMM), Support Vector Machine (SVM), and Conditional random field (CRF). The major ingredients of the deep learning based approaches are LSTM, BLSTM, and RNN. In this work, the proposed PoS tagger is also

TABLE 6. Comparison of RNN and LSTM-based models with and without word embedding weights.

Models	Description	Training %	Test %
RNN with non embedding weights	64	93.80	93.45
RNN with word embedding weights	64	98.86	98.76
BLSTM Model with non-embedding weights	64	93.00	92.77
BLSTM Model with word embedding weights	64	99.74	98.82

based on LSTM architectures. This work also validates the effectiveness of deep learning based approaches as reported in [49]. In terms of novelty, our model uses the word embedding weights in the initial layers. To the best of our knowledge, it is the first time that we use the fastext corpus that presents word embedding vector regarding the Pashto language. It is empirically shown that when including the word embedding weights, the accuracy improves by minimum of 5

Usually, the PoS taggers are suffered due to similarity and ambiguity in similar words. And such ambiguity can only be understood while looking into the context. Therefore, the proposed tagset has all its constituent words in their valid composed structure that is abide by the grammatical rules of the Pashto language. In short, we did not alter its natural flow, and hence preserved the very important aspect of context. As a result when added with word embedding vectors the performance has improved. Despite of these good stories, the conjunction 'CCONJ', ad-position 'ADP', verb 'VERB' and pronoun 'PRON' are the PoS tags with highest miss classification. In future, we will work with approaches that could improve the accuracy regarding the miss classified PoS tags.

VI. CONCLUSION AND FUTURE WORK

The PoS tagging in the Pashto language using a deep learning approach is presented for the first time in this research paper. This study used deep learning techniques on a sizable Pashto dataset to specifically observe the Part of Speech tagging in Pashto language.

Additionally, the research takes a step further and employs a deep learning LSTM model to ascertain other ways to use part-of-speech tagging in Pashto. We used the more effective deep learning LSTM model, and by using this model, we attained 98.8% accuracy. This work is the first to use a deep learning approach for the Pashto language, use a big dataset, and achieve high accuracy.

Part of Speech tagging is essential to create effective Natural Language Processing (NLP) parsers [50] and applications for Pashto. We can generate better results by expanding the corpus and adding additional tags to the tagset. Future work will focus on using deep learning to construct a parser for the Pashto language.

ACKNOWLEDGMENT

The authors would like to thank Prof. Ali Khail Daryab, Chairperson, Department of Pashto Language, University of Malakand, and Prof. Nasrallah Jan Wazir, Director, Pashto Academy, University of Peshawar, for their help and support regarding Pashto text collection.

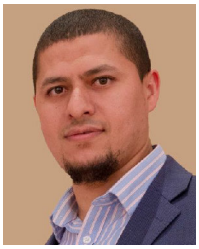
REFERENCES

[1] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Comput. Linguistics*, vol. 21, no. 4, pp. 543–565, 1995.
[2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," 2020, *arXiv:2004.12158*.
[3] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Comput. Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.
[4] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, and T. Breuel, "Scale and rotation invariant OCR for pashto cursive script using MDLSTM network," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1101–1105.
[5] H. Khalil, "Pashtoon culture in Pashto Tappa," Ph.D. dissertation, Nat. Inst. Historical Cultural Res., Quaid-I-Azan, Univ. Islamabad, Pakistan, 2011.
[6] X. Zheng, H. Chen, and T. Xu, "Deep learning for Chinese word segmentation and POS tagging," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 647–657.
[7] J. Kupiec, "Robust part-of-speech tagging using a hidden Markov model," *Comput. Speech Lang.*, vol. 6, no. 3, pp. 225–242, Jul. 1992.
[8] T. Horsmann and T. Zesch, "Do LSTMs really work so well for PoS tagging?—A replication study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 727–736.
[9] C. Ding, H. T. Z. Aye, W. P. Pa, K. T. Nwet, K. M. Soe, M. Utiyama, and E. Sumita, "Towards burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 1, pp. 1–34, Jan. 2020.
[10] I. Hussain, R. Ahmad, S. Muhammad, K. Ullah, H. Shah, and A. Namoun, "PHTI: Pashto handwritten text imagebase for deep learning applications," *IEEE Access*, vol. 10, pp. 113149–113157, 2022.
[11] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, pp. 64–67, Dec. 2001.
[12] B. Plank, A. Søgaard, and Y. Goldberg, "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss," 2016, *arXiv:1604.05529*.
[13] M. Okhovvat and B. M. Bidgoli, "A hidden Markov model for Persian part-of-speech tagging," *Proc. Comput. Sci.*, vol. 3, pp. 977–981, 2011.
[14] F. M. Hasan, "Comparison of different POS tagging techniques for some south Asian languages," Ph.D. dissertation, Dept. Comput. Sci. Eng., BRAC Univ., Dhaka, Bangladesh, 2006.
[15] S. Mohammed, "Using machine learning to build POS tagger for under-resourced language: The case of somali," *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 717–729, Sep. 2020.
[16] K. Darwish, H. Mubarak, M. Eldesouki, A. Abdelali, Y. Samih, R. Alharbi, M. Attia, W. Magdy, and L. Kallmeyer, "Multi-dialect Arabic POS tagging: A CRF approach," in *Proc. 11th, Ed., Lang. Resour. Eval. Conf.*, Paris, France: European Language Resources Association, 2018, pp. 93–98.
[17] V. Makarenkov, L. Rokach, and B. Shapira, "Choosing the right word: Using bidirectional LSTM tagger for writing support systems," *Eng. Appl. Artif. Intell.*, vol. 84, pp. 1–10, Sep. 2019.
[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[19] L. Marquez, L. Padro, and H. Rodriguez, "A machine learning approach to POS tagging," *Mach. Learn.*, vol. 39, pp. 59–91, Jun. 2000.

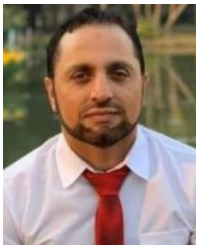
[20] H. Amiri, F. Raja, M. Sarmadi, S. Tasharofi, H. Hojjat, and F. Oroumchian, "A survey of part of speech tagging in Persian," *Data Base Res. Group*, Jan. 2007.
[21] A. Baig, M. U. Rahman, H. Kazi, and A. Baloch, "Developing a POS tagged corpus of Urdu tweets," *Computers*, vol. 9, no. 4, p. 90, Nov. 2020.
[22] K. Alrajhi et. al., "Automatic Arabic part-of-speech tagging: Deep learning neural LSTM versus Word2 Vec," *Int. J. Comput. Digit. Syst.*, vol. 8, no. 3, pp. 307–315, Jul. 2019.
[23] I. Rabbi, M. A. Khan, and R. Ali, "Rule-based part of speech tagging for Pashto language," in *Proc. Conf. Lang. Technol.*, Lahore, Pakistan, 2009.
[24] K. K. Akhil, R. Rajimol, and V. S. Anoop, "Parts-of-speech tagging for Malayalam using deep learning techniques," *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 741–748, Sep. 2020.
[25] G. Leech, "Corpus annotation schemes," *Literary Linguistic Comput.*, vol. 8, no. 4, pp. 275–281, Oct. 1993.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[27] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," 2017, *arXiv:1801.01078*.
[28] Y. Luan, Y. Ji, and M. Ostendorf, "LSTM based conversation models," 2016, *arXiv:1603.09457*.
[29] J. H. Wells and L. R. Williams, *Embeddings and Extensions in Analysis*, vol. 84. Berlin, Germany: Springer, 2012.
[30] M. H. Sazli, "A brief review of feed-forward neural networks," *Commun. Fac. Sci. Univ. Ankara Ser. A2-A3 Phys. Sci. Eng.*, vol. 50, no. 1, pp. 11–17, 2006.
[31] W. Fang, Y. Chen, and Q. Xue, "Survey on research of RNN-based spatio-temporal sequence prediction algorithms," *J. Big Data*, vol. 3, no. 3, pp. 97–110, 2021.
[32] G. Prabha, P. V. Jyothsna, K. K. Shahina, B. Premjith, and K. P. Soman, "A deep learning approach for part-of-speech tagging in nepali language," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 1132–1136.
[33] S. Hochreiter. (1991). *Untersuchungen Zu Dynamischen Neuronalen Netzen. 1991*. [Online]. Available: http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf
[34] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.
[35] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative LSTM for sequence modelling," 2016, *arXiv:1609.07959*.
[36] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.
[37] Y. Li and T. Yang, "Word embedding for understanding natural language: A survey," in *Guide to Big Data Applications*. Cham, Switzerland: Springer, 2018, pp. 83–104.
[38] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Syst.*, vol. 151, pp. 78–94, Jul. 2018.
[39] X. Wang, D. Bo, C. Shi, S. Fan, Y. Ye, and P. S. Yu, "A survey on heterogeneous graph embedding: Methods, techniques, applications and sources," *IEEE Trans. Big Data*, vol. 9, no. 2, pp. 415–436, Apr. 2023.
[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[41] S. T. Gries, "What is corpus linguistics?" *Lang. Linguistics Compass*, vol. 3, no. 5, pp. 1225–1241, 2009.
[42] L. Shen and M. Xu, "Student public opinion management in campus commentary based on deep learning," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–12, Apr. 2022.
[43] D. Nadeau and P. D. Turney, "A supervised learning approach to acronym identification," in *Proc. Conf. Canadian Soc. Comput. Stud. Intell.* Cham, Switzerland: Springer, 2005, pp. 319–329.
[44] Y. Xie, C. Li, B. Yu, C. Zhang, and Z. Tang, "A survey on dynamic network embedding," 2020, *arXiv:2006.08093*.
[45] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses," in *Proc. Int. Joint Conf. Neural Netw.*, 1989, pp. 191–196.
[46] N. K. S. Kumar and N. Malarvizhi, "Bi-directional LSTM–CNN combined method for sentiment analysis in part of speech tagging (PoS)," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 373–380, Jun. 2020.
[47] S. Wang, W. Zhou, and C. Jiang, "A survey of word embeddings based on deep learning," *Computing*, vol. 102, no. 3, pp. 717–740, Mar. 2020.
[48] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C.-J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, 2019.
[49] A. Chiche and B. Yitagesu, "Part of speech tagging: A systematic review of deep learning and machine learning approaches," *J. Big Data*, vol. 9, no. 1, pp. 1–25, Jan. 2022.
[50] E. M. Sibarani, M. Nadial, E. Panggabean, and S. Meryana, "A study of parsing process on natural language processing in bahasa Indonesia," in *Proc. IEEE 16th Int. Conf. Comput. Sci. Eng.*, Dec. 2013, pp. 309–316.

SHAHEEN ULLAH received the master's degree from the Department of Computer Science, University of Peshawar, Pakistan, and the M.S. degree in computer science from the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal. He is currently serving as a Lecturer at Government Degree College, Dir Upper.

RIAZ AHMAD received the M.S. degree (Hons.) in computer science from NUCES (FAST) University, Pakistan, in 2010, and the Ph.D. degree from the Technical University of Kaiserslautern, Germany, in 2018. He was a member of the Multimedia Analysis and Data Mining (MADM) Research Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. He is currently heading the Computer Science Department, Shaheed Benazir Bhutto University, Sheringal, Pakistan. His research interests include document image analysis, image processing, and optical character recognition. More specifically, his work examines the challenges posed by cursive script languages in the field of OCR systems. In addition to that, he is studying the behavior of deep learning architectures in the field of OCR in terms of invariant approaches against scale and rotation variation in Pashto cursive text.



ABDALLAH NAMOUN (Member, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in informatics from The University of Manchester, U.K., in 2004 and 2009, respectively. He is currently a Full Professor of intelligent interactive systems and the Head of Information Technology Department, Faculty of Computer and Information Systems, Islamic University of Madinah. He has authored more than 90 publications in research areas spanning intelligent systems, human–computer interaction, machine learning, smart cities, and technology acceptance. He has extensive experience in leading complex research projects (worth more than 23 million Euros) with several distinguished SMEs, such as SAP, BT, and ATOS. His recent research interests focus on integrating state of the art artificial intelligence models in the development of interactive systems and smart spaces.



SIRAJ MUHAMMAD received the M.Phil. degree from Quaid-i-Azam University, Islamabad, Pakistan, in 2010, and the Ph.D. degree from the Asian Institute of Technology (AIT), Thailand, in 2020. He was a Software Engineer at Elixir Technologies, Islamabad, from 2010 to 2011. He is currently an Assistant Professor with the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Pakistan. His research interests include reverse engineering, computer vision, image processing, deep learning, and natural language processing.

KHALIL ULLAH received the degree in computer systems engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2006, the Master of Science (M.S.) degree in electronics and communications engineering from Myongji University, South Korea, in 2009, and the Ph.D. degree in biomedical engineering from LISiN, Politecnico di Torino, in 2016, under the Erasmus Mundus Expert II Fellowship. He is currently an Assistant Professor and the Head of the Software Engineering Department, University of Malakand. His research interests include extracting muscle anatomical and physiological information from high-density electromyography, computer vision, digital signal and image processing, and deep learning with applications to medical healthcare.

IBRAR HUSSAIN received the B.C.S. degree (Hons.) from the Department of Computer Science, University of Peshawar, Pakistan, and the M.S. degree in computer science from the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Technology, University of Malakand, Khyber Pakhtunkhwa, Pakistan.

ISA ALI IBRAHIM is a FCIIS, FBCS, FNCS, and the Chancellor of Public University, Sudan, and a Governing Council member of two universities. He has been also a Full Professor of cybersecurity with the Federal University of Technology, Owerri, since March 2021, where he teaches probono. He served as the Chief Digital Officer of Nigeria as well as the Cabinet Minister of Communications and Digital Economy of Nigeria, from 2019 to 2023. He was appointed by the United Nations through ITU to serve as the Chairperson of over 194+ global ministers of ICT through WSIS 2022. He authored over 10 books on emerging technologies, among others, published over 35 journal articles, as well as attended over 190 conferences as the keynote speaker, presenters, or discussants in over 40 countries globally.