[1] Afsana Laskar
[2] Shikhar Kumar Sarma
[3] Jessica Saikia
[3] Dikshita Borah

A Contextual and Embedded Approach for Part-of-Speech Tagging for Assamese- English Code-Mixed Text

Journal of Electrical Systems

Abstract: - Part-of-speech (POS) tagging is an essential procedure in natural language processing (NLP) that allocates each word in a text to its appropriate grammatical category, including nouns, verbs, adjectives, and others. Part-of-speech tagging in code-mixed texts presents challenges, because of language mixing and the lack of large annotated datasets, especially in low-resource languages like Assamese. This paper presents a comparative analysis of different POS tagging models to develop a hybrid system for POS tagging of Assamese-English code-mixed texts. Each model's performance is evaluated on Assamese-English code-mixed dataset, analysing metrics like accuracy, precision, and recall. Based on these findings, we propose a conceptual and embedded POS tagging system that combines the strengths of XLM-RoBERTa model and CRF model to enhance overall tagging accuracy.

Keywords: Assamese-English, POS tagging, HMM, LSTM, XLM-RoBERTa, CRF

I. INTRODUCTION

With the rise of multilingualism in online communication, code-mixing, where speakers use multiple different languages within a single conversation, has become a common phenomenon, particularly in regions like India where multilingual communities thrive.

Assamese-English code-mixed texts, frequently observed on social media platforms, present unique challenges. PoS tagging is one such NLP task that helps in syntactic parsing, machine translation, and other higher-level NLP applications like Assamese [2].

The task of Part-of-Speech tagging in code-mixed texts is simply more challenging because of various aspects, including the necessity to ascertain the language of each word, address grammatical discrepancies between the languages, and manage out-of-vocabulary terms. Assamese, being a morphologically rich but underrepresented language in the NLP domain, lacks sufficient annotated corpora, making it even more challenging to apply standard machine-learning techniques that require large datasets for training.

The proposed contextual and embedded approach is evaluated on a newly created Assamese-English code-mixed corpus. Results show that this method outperforms baseline models, highlighting the potential of hybrid approaches for PoS tagging in code-mixed environments. This research adds to the expanding domain of NLP for low-resource languages and demonstrates the importance of specialized strategies for multilingual and code-mixed text.

II. LITERATURE REVIEW

PoS tagging has been a long fundamental task in NLP, with early rule-based methods facing scalability issues for complex and morphologically rich languages. Statistical models like Conditional Random Fields (CRF)[10] and Hidden Markov Models (HMM) offered more flexibility but struggled to generalize for code -mixed or low resource language.

Code-mixed texts introduce additional challenges, such as language identification and unpredictable grammatical structures, which hinder the effectiveness of conventional PoS taggers trained on monolingual corpora. To address these issues, hybrid models combining rule-based and statistical approaches have gained attention. For instance, Phukan et al. [1] used LSTM and Bi-LSTM models, achieving accuracies of 92.80% and 93.36%, respectively, with deep learning techniques improving tagging performance. Talukdar et al. [2] explored Assamese PoS tagging

[1] *Corresponding author: Department of Information Technology, Gauhati University,Gauhati, Assam,Email:laskar.afs@gmail.com
[2] Author 2 : Department of Information Technology, Gauhati University, Gauhati, Assam, Email: sks001@gmail.com
[3] Author 3: The Assam Royal Global University,Gauhati,Assam, Email:jessicasaikia8@gmail.com
[3] Author 3: The Assam Royal Global University,Gauhati,Assam, Email:dikshitaborah24@gmail.com

using RNN and GRU, transitioning to the UPoS framework. The GRU model achieved 94.38% accuracy, outperforming the RNN model. This study was the first to apply UPoS tagging for Assamese with deep learning, setting a baseline for future work. Dowlagar et al.[4] proposed a joint model that integrates Language Identification and PoS tagging using BERT combined with CNNs, improving multilingual analysis by utilizing interdependencies between these tasks The paper by Talukdar, Kuwali, et al.[5] proposes a deep learning-based approach for Universal Part-of-Speech (UPoS) tagging of Assamese religious texts. It employs a BiLSTM-CRF model, leveraging pre-trained word embeddings specific to the Assamese language for effective tagging. Research on PoS tagging for Assamese includes Talukdar, Kuwali, et.al[6], paper focuses on converting PoS tags into Universal Part-of-Speech (UPoS) tags for the Assamese language to align with universal linguistic standards. The authors created UPoS-tagged resources for Assamese by mapping existing PoS annotations to the UPoS scheme, addressing challenges like linguistic nuances and non-standard scripts. This work contributes to improving the compatibility of Assamese linguistic resources with multilingual NLP frameworks and facilitates cross-lingual analysis.

III. METHODOLOGY

A. Dataset

For this research, a corpus containing 1,00,627 Assamese-English code- English words or phrases was inserted manually to reflect code-mixed patterns. The dataset is formatted in a .csv file, with each sentence paired with its corresponding POS tags.

```
1   Sentence,POS_Tags
2   আমি love the অসমীয়া folk সঙ্গীত ,AS-PRON EN-VERB EN-DT AS-NOUN EN-NOUN AS-NOUN
3   Such a beautiful বন made my day,EN-PRON EN-DT EN-NOUN AS-NOUN EN-VERB EN-PRON EN-NOUN
4   আমি তোমাৰ সৈতে spend কৰা time is always মজা,AS-PRON AS-PRON AS-ADV EN-VERB AS-VERB EN-NOUN EN-VERB EN-ADV AS-ADJ
5   The বতাহ left a beautiful impression,EN-DT AS-NOUN EN-NOUN EN-DT EN-NOUN EN-NOUN
6   Such a amazing ৰং ruined my day,EN-PRON EN-DT EN-VERB AS-NOUN EN-VERB EN-PRON EN-NOUN
7   I felt shoddy about the আকাশ,EN-PRON EN-NOUN EN-NOUN EN-ADJ EN-DT AS-NOUN
8   I had a lovely experience with the চহৰ,EN-PRON EN-VERB EN-DT EN-ADJ EN-NOUN EN-PREP EN-DT AS-NOUN
9   The ৰেষ্টুৰেন্ট was disappointing which made me happy,EN-DT AS-NOUN EN-VERB EN-VERB EN-PRON EN-VERB EN-PRON EN-ADJ
10  The পৰিয়াল is pathetic ,EN-DT AS-NOUN EN-VERB EN-ADJ
```

Fig 1: Dataset

B. Data Annotation

The dataset was carefully annotated by assigning each word (or token) a specific part-of-speech (POS) tag to prepare it for training the models. Additionally, custom tags were added to better handle unique linguistic characteristics found in code-mixed Assamese-English text, which, helped identify and differentiate specific language patterns this manual annotation process ensured high accuracy and allowed the model to better understand language patterns and dependencies during training, ultimately improving its POS tagging performance. Custom Part of Speech tags used are: 1. Noun: EN-NOUN, AS-NOUN 2. Pronoun: EN-PRON,AS-PRON 3.Verb:EN-VERB, AS-VERB, 4.Adverb:EN-ADV,AS-ADV 5.Preposition:EN-PREP, AS-PREP 6. Adjective: EN-ADJ, AS-ADJ 7. Conjuction:EN-CONJ, AS-CONJ 8. Interjection: EN-INTJ 9. Determiners: EN-DT.

C. Training and Testing:

Initially, the individual models were trained and tested on gradually increasing the subsets of the dataset, specifically 20,000, 50,000, 75,000, and ultimately 1,00,627 code-mixed sentences, while tracking accuracy at each stage. The dataset had been split into training and testing sets of ratios of 80:20 and 70:30. Following the evaluation of individual models, combinations of models were implemented and tested. The hybrid models were assessed by employing same metrics as the individual models, focusing on their ability to handle transitions between Assamese and English words and resolve ambiguities in code-mixed words. The hybrid model's performance was further compared to the baseline results of the individual models to measure improvements and effectiveness.

IV. COMPARATIVE EVALUATION OF DIFFERENT MODELS

In this section, the comparative analysis of individual models followed by their performance evaluation, based on precision, accuracy, and recall, is discussed. The individual models tested were: Rule-based model, HMM (Hidden Markow Model), CRF model, LSTM (Long Short-Term Memory), BiLSTM (Bidirectional Long Short-Term Memory), mBERT, XLM- RoBERTa, and MuRIL. After the successful building, training, and evaluation of the

models are done, the focus is on combining the strengths of individual models to form various combinations for a hybrid model. The combination models tested were: CRF+BiLSTM, XLM-RoBERTa+BiLSTM, and XLM-RoBERTa+CRF. The same evaluation metrics are considered to analyze these models.

A. Rule Based Method:

The evaluation of models for POS tagging in English-Assamese code-mixed text covered a diverse range of approaches, starting with the Rule-based model, which relies on predefined linguistic rules to predict POS tags. While it performs well on predictable patterns, it struggles with linguistic variability in code-mixed text. These rules are based on syntax, word morphology, and contextual patterns. The model analyses the surrounding context of each word to determine its tag, following these manually defined patterns.For example, a word ending in '-ing' might be tagged as a verb if it follows a noun, or as a noun if it appears in a phrase like "the running." The model analyses the sentence structure and applies these rules to determine the most appropriate POS tag for each word. While effective in structured contexts, this approach depends heavily on the quality and coverage of the rules, which must be manually defined and then fine-tuned.

B. Hidden Markow Model:

The Hidden Markov Model (HMM), a probabilistic sequence model, demonstrated moderate success by using state transitions, but it lacked the capacity to capture long-range dependencies.

C. Conditional Random Field :

The Conditional Random Field (CRF) model improved upon HMM by incorporating global sequence optimization and feature engineering, but it still required manually crafted features for best performance.

*D. Long Short Term Memory:*The LSTM network, capable of learning long-range dependencies, outperformed traditional models by learning contextual representations from the text.

E. BiLSTM: The BiLSTM (Bidirectional LSTM) improved this by analyzing the sequence in both forward and backward directions, facilitating a more comprehensive context acquisition.

F. For transformer-based models, mBERT, a multilingual variant of BERT, showed strong contextual understanding but had limitations in handling underrepresented languages like Assamese.

G. XLM-RoBERTa, pre-trained on a diverse multilingual corpus, excelled in capturing both syntactic and semantic information in code-mixed text.

H. Similarly, MuRIL, fine-tuned for Indian languages, demonstrated strong performance but was less effective than XLM-RoBERTa for code-mixed tasks.

I. Hybrid Models: Hybrid models combining neural and probabilistic approaches were also evaluated. CRF+BiLSTM combined BiLSTM's contextual embeddings with CRF's sequence optimization, resulting in improved tag predictions. XLM-RoBERTa+BiLSTM combined transformer-based embeddings with BiLSTM, adding sequential processing but lagging slightly behind in global sequence optimization. The XLM-RoBERTa+CRF model emerged as the most effective, integrating deep multilingual contextual embeddings with CRF's ability to optimize label sequences, making it specifically appropriate for addressing the complexities of POS tagging in code-mixed text.

V. RESULTS AND DISCUSSION

The models were assessed utilizing precision, recall, and accuracy as primary measures. The dataset was uniformly divided into training and testing sets utilizing 80:20 and 70:30 ratios for all models. For the rule-based model, a bilingual Assamese- English dictionary was used alongside the dataset, which was initially split 80:20 for practice and later 70:30 for training and testing. The HMM attained an accuracy of 91%, while the Conditional Random Fields (CRF) model, which utilized feature data and labels, attained an accuracy of 94%. The LSTM model was trained with word indices and POS tag labels, where training continued for a set number of epochs to prevent overfitting. A 93% accuracy was achieved for LSTM model. For the BiLSTM, tokenised sentences were used as input with POS tags as ground truth, and the model processed 32 sentences per batch. After 10 epochs, it achieved an accuracy of 95%. The mBERT model was trained using a linear warmup schedule followed by decay to stabilize

the training process. Cross-entropy loss was employed for token classification. Following 20 epochs of training with a batch size of 16, the model attained an accuracy of 88%. The XLM-RoBERTa model utilized the Hugging Face Tokenizer to transform phrases into token IDs suitable for the model, and training was performed with a batch size of 32 across 40 epochs. This particular model attained an accuracy of 94%.

The MuRIL model (Multilingual Representations for Indian Languages) was evaluated as part of the study. It was trained to process batches of 32 sentences simultaneously, with a batch size 32, over a total of 30 training epochs.

The training used two learning rates, 1×10^{-5} and 2×10^{-5}, to fine-tune the model's weights. These learning rates helped control how much the model updated its parameters during training, ensuring it learned effectively without overfitting or diverging.

The Adam optimizer was employed to reduce the Cross-Entropy Loss, a prevalent loss function for classification problems. This loss function quantifies the disparity between expected and actual class probabilities, directing the model to enhance its predictions progressively.

After completing the training, the MuRIL model achieved an impressive accuracy of 96%, indicating its high effectiveness in handling the classification tasks within the given dataset.

The hybrid models were then evaluated. The CRF+BiLSTM combination utilized an 80:20 dataset split, resulting in an accuracy of 92%. The XLM-RoBERTa+BiLSTM model had been trained by employing a learning rate of 5e-5, a batch size of 32, and for 20 epochs, processing tokenized sentences with a maximum length of 128 and a BiLSTM hidden dimension of 128. The accuracy achieved for this combination is 93%. Lastly, for the XLM-RoBERTa+CRF hybrid, the dataset was split into 80:20 and 70:30 ratios, and the model was trained using the AdamW optimiser, with a learning rate of 5e-5 for XLM-RoBERTa and 1e-3 for the CRF layer. This combination achieved the highest accuracy of 97%. The XLM-RoBERTa + CRF hybrid model emerged as the best-performing model due to its ability to handle the complexities of English-Assamese code-mixed text effectively. XLM-RoBERTa can effectively identify rich semantic and syntactic information, making it highly adept at understanding linguistic complexities and context in code-mixed sentences. The addition of a Conditional Random Field (CRF) layer further enhances performance by modelling dependencies between output labels, such as the likelihood of specific tags following others, which is crucial in POS tagging tasks. Architecturally, the model processes tokenised code-mixed sentences through XLM-RoBERTa to generate deeply contextualised embeddings, which are then passed through a feature projection layer before being fed into the CRF. The CRF utilizes a transition matrix to predict the optimal sequence of tags based on contextual dependencies. This hybrid methodology integrates the contextual efficacy of XLM-RoBERTa with the structural prediction proficiency of CRF, resulting in enhanced performance regarding precision, recall, and F1 score relative to alternative configurations. Figure 2 displays classification report of HMM model. Figure 3 displays classification report of CRF model. Figure 4 displays classification report of LSTM model and Figure 5 displays classification report of BiLSTM model. Figure 6 displays the classification report of mBERT model and Figure 7 XLM-RoBERTa model classification report and Figure 8 represnts MuRIL model classification report. Figure 9 represents the hybrid models CRF+BiLSTM model, Figure 10 XLM-RoBERTa+BiLSTM model, Figure 11 represent XLM-RoBERTa+CRF. Figure 12 displays the evaluation metrics for every model assessed and applied in this research.

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.91	0.95	0.93	16473
EN-NOUN	0.70	0.73	0.71	22238
EN-VERB	0.80	0.72	0.76	19226
AS-NOUN	0.81	0.73	0.77	16413
EN-PRON	1.00	0.98	0.99	10079
AS-VERB	0.71	0.73	0.72	2761
EN-PREP	0.96	0.91	0.93	2402
AS-ADJ	0.81	0.87	0.84	3869
EN-ADJ	0.97	0.84	0.90	4736
EN-ADV	0.95	0.89	0.92	1327
EN-CONJ	0.93	0.78	0.85	1068
AS-ADV	0.91	0.80	0.85	199
AS-CONJ	0.79	0.74	0.76	151
AS-PRON	0.96	0.94	0.95	459
accuracy			0.91	101401
weighted avg	0.83	0.81	0.82	101401

Fig. 2 HMM model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.80	0.84	0.82	16473
EN-NOUN	0.94	0.76	0.84	22238
EN-VERB	0.79	0.83	0.81	19226
AS-NOUN	0.79	0.72	0.75	16413
EN-PRON	0.87	0.92	0.89	10079
AS-VERB	0.87	0.97	0.92	2761
EN-PREP	0.73	0.75	0.74	2402
AS-ADJ	0.86	0.96	0.91	3869
EN-ADJ	0.76	0.74	0.75	4736
EN-ADV	0.78	0.86	0.82	1327
EN-CONJ	0.99	0.85	0.91	1068
AS-ADV	0.88	0.75	0.81	199
AS-CONJ	0.94	0.91	0.92	151
AS-PRON	1.00	0.92	0.96	459
accuracy			0.94	101401
weighted avg	0.84	0.81	0.82	101401

Fig 3. CRF model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.95	0.79	0.86	16473
EN-NOUN	0.91	0.98	0.94	22238
EN-VERB	0.92	0.92	0.92	19226
AS-NOUN	0.93	0.92	0.92	16413
EN-PRON	0.78	0.90	0.84	10079
AS-VERB	0.76	0.98	0.86	2761
EN-PREP	0.93	0.88	0.90	2402
AS-ADJ	0.95	0.96	0.95	3869
EN-ADJ	0.73	0.91	0.81	4736
EN-ADV	0.96	0.87	0.91	1327
EN-CONJ	0.80	0.88	0.84	1068
AS-ADV	0.87	0.83	0.85	199
AS-CONJ	0.78	1.00	0.88	151
AS-PRON	0.88	0.85	0.86	459
accuracy			0.93	101401
weighted avg	0.90	0.91	0.90	101401

Fig. 4 LSTM model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.98	0.73	0.84	16473
EN-NOUN	0.93	0.80	0.86	22238
EN-VERB	0.87	0.79	0.83	19226
AS-NOUN	0.95	0.91	0.93	16413
EN-PRON	0.92	1.00	0.96	10079
AS-VERB	0.86	0.87	0.86	2761
EN-PREP	0.78	0.74	0.76	2402
AS-ADJ	0.91	0.70	0.79	3869
EN-ADJ	0.92	0.75	0.83	4736
EN-ADV	0.89	0.77	0.83	1327
EN-CONJ	0.92	0.85	0.88	1068
AS-ADV	0.84	0.71	0.77	199
AS-CONJ	0.78	0.70	0.74	151
AS-PRON	0.89	0.90	0.89	459
accuracy			0.95	101401
weighted avg	0.92	0.82	0.87	101401

Fig 5. BiLSTM model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.83	0.83	0.83	16473
EN-NOUN	0.99	0.91	0.95	22238
EN-VERB	0.90	0.76	0.82	19226
AS-NOUN	0.92	0.73	0.81	16413
EN-PRON	0.86	0.99	0.92	10079
AS-VERB	0.74	0.74	0.74	2761
EN-PREP	0.88	0.77	0.82	2402
AS-ADJ	0.71	0.83	0.77	3869
EN-ADJ	0.78	0.79	0.78	4736
EN-ADV	0.99	0.80	0.88	1327
EN-CONJ	0.71	0.95	0.81	1068
AS-ADV	0.94	0.84	0.89	199
AS-CONJ	0.98	0.77	0.86	151
AS-PRON	0.94	0.72	0.82	459
accuracy			0.88	101401
weighted avg	0.89	0.83	0.85	101401

Fig 6. mBERT model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.85	0.90	0.87	16473
EN-NOUN	0.88	0.96	0.92	22238
EN-VERB	0.89	0.93	0.91	19226
AS-NOUN	0.87	0.81	0.84	16413
EN-PRON	0.74	0.76	0.75	10079
AS-VERB	0.86	0.92	0.89	2761
EN-PREP	0.76	0.73	0.74	2402
AS-ADJ	0.88	0.79	0.83	3869
EN-ADJ	0.84	0.85	0.84	4736
EN-ADV	0.88	0.75	0.81	1327
EN-CONJ	0.98	0.87	0.92	1068
AS-ADV	0.81	0.88	0.84	199
AS-CONJ	0.72	0.73	0.72	151
AS-PRON	0.94	0.90	0.92	459
accuracy			0.94	101401
weighted avg	0.86	0.88	0.87	101401

Fig 7: *XLM-RoBERTa model classification report*

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.82	0.75	0.78	16473
EN-NOUN	0.84	0.88	0.86	22238
EN-VERB	0.74	0.92	0.82	19226
AS-NOUN	0.75	0.76	0.75	16413
EN-PRON	0.83	0.84	0.83	10079
AS-VERB	0.81	0.88	0.84	2761
EN-PREP	0.99	0.98	0.98	2402
AS-ADJ	0.92	0.95	0.93	3869
EN-ADJ	0.77	0.94	0.85	4736
EN-ADV	0.77	0.99	0.87	1327
EN-CONJ	0.84	0.98	0.90	1068
AS-ADV	0.91	0.86	0.88	199
AS-CONJ	0.97	0.89	0.93	151
AS-PRON	0.78	0.86	0.82	459
accuracy			0.96	101401
weighted avg	0.80	0.85	0.82	101401

Fig 8: MuRIL model classification report

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.72	0.73	0.72	16473
EN-NOUN	0.99	0.79	0.88	22238
EN-VERB	0.72	0.81	0.76	19226
AS-NOUN	0.77	0.73	0.75	16413
EN-PRON	0.74	0.91	0.82	10079
AS-VERB	0.79	0.85	0.82	2761
EN-PREP	0.71	0.78	0.74	2402
AS-ADJ	0.92	0.86	0.89	3869
EN-ADJ	0.98	0.95	0.96	4736
EN-ADV	0.73	0.75	0.74	1327
EN-CONJ	0.84	0.84	0.84	1068
AS-ADV	0.75	0.86	0.80	199
AS-CONJ	0.82	0.82	0.82	151
AS-PRON	0.98	0.72	0.83	459
accuracy			0.92	101401
weighted avg	0.81	0.80	0.80	101401

Fig 9:CRF+BiLSTM model

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.89	0.87	0.88	16473
EN-NOUN	0.72	0.73	0.72	22238
EN-VERB	0.91	0.78	0.84	19226
AS-NOUN	0.79	0.86	0.82	16413
EN-PRON	0.83	0.95	0.89	10079
AS-VERB	0.79	0.86	0.82	2761
EN-PREP	0.79	0.90	0.84	2402
AS-ADJ	0.96	0.91	0.93	3869
EN-ADJ	0.76	0.85	0.80	4736
EN-ADV	0.90	0.92	0.91	1327
EN-CONJ	0.78	0.77	0.77	1068
AS-ADV	0.92	0.73	0.81	199
AS-CONJ	0.84	0.86	0.85	151
AS-PRON	0.73	0.79	0.76	459
accuracy			0.93	101401
weighted avg	0.82	0.83	0.82	101401

Fig.10 XLM-RoBERTa+BiLSTM model

Classification Report:

	precision	recall	f1-score	support
EN-DT	0.74	0.88	0.80	16473
EN-NOUN	0.72	0.80	0.76	22238
EN-VERB	0.83	0.97	0.89	19226
AS-NOUN	0.78	0.76	0.77	16413
EN-PRON	0.86	0.80	0.83	10079
AS-VERB	0.81	0.99	0.89	2761
EN-PREP	0.80	0.71	0.75	2402
AS-ADJ	0.98	0.88	0.93	3869
EN-ADJ	0.72	0.75	0.73	4736
EN-ADV	0.99	0.84	0.91	1327
EN-CONJ	0.89	0.84	0.86	1068
AS-ADV	0.78	0.76	0.77	199
AS-CONJ	0.80	0.76	0.78	151
AS-PRON	0.82	0.75	0.78	459
accuracy			0.97	101401
weighted avg	0.79	0.84	0.81	101401

Fig.11 XLM-RoBERTa+CRF model

Model	Precision	Recall	F1 score	Accuracy
HMM	83%	81%	82%	91%
CRF	84%	81%	82%	94%
LSTM	90%	91%	90%	93%
BiLSTM	92%	82%	87%	95%
mBERT	89%	83%	85%	88%
XLM-RoBERTa	86%	88%	87%	94%
MuRIL	80%	85%	82%	96%
CRF+BiLSTM	81%	80%	80%	92%
XLM-RoBERTa + BiLSTM	82%	83%	82%	93%
XLM-RoBERTa + CRF	79%	84%	81%	97%

Fig 12. Performance metrics of different models

VI. CONCLUSION

The current study presents a hybrid approach to POS tag Assamese-English code-mixed texts, by combining the XLM-RoBERTa model with Conditional Random Fields (CRF), utilizing our own annotated dataset, manually curating a set of sentences for training and evaluating the model effectively. The performance of our hybrid model demonstrated promising results, achieving a 97% accuracy on the testing dataset. This performance is competitive with existing approaches in code-mixed PoS tagging. Looking ahead, further research can be done on expanding the model to other languages and addressing challenges such as word- level transliteration and out-of-vocabulary terms. Exploring domain-specific corpora or leveraging other sequence models could further enhance the model's generalization and accuracy across diverse code-mixed scenarios.

REFERENCES

[1] Phukan, Rituraj, et al. "Exploring Character-Level Deep Learning Models for POS Tagging in Assamese Language." Procedia Computer Science 235 (2024): 1467-1476.

[2] Talukdar, Kuwali, and Shikhar Kumar Sarma. "Deep Learning based Part-of-Speech tagging for Assamese using RNN and GRU." Procedia Computer Science 235 (2024): 1707-1712.

[3] Pradhan, Ashish, and Archit Yajnik. "Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM." Multimedia Tools and Applications 83.4 (2024): 9893-9909.

[4] Suman Dowlagar and Radhika Mamidi. 2021. A Pre-trained Transformer and CNN Model with Joint Language ID and Part-of-Speech Tagging for Code-Mixed Social-Media Text. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 367–374, Held Online. INCOMA Ltd.

[5] Talukdar, Kuwali, et al. "Deep Learning based UPoS Tagger for Assamese Religious Text." International Journal of Religion 5.4 (2024): 163-170.

[6] Talukdar, Kuwali, and Shikhar Kumar Sarma. "PoS to UPoS Conversion and Creation of UPoS Tagged Resources for Assamese Language." In Proceedings of the 20th International Conference on Natural Language Processing (ICON), pp. 450-459. 2023.

[7] Pathak, Dhrubajyoti, et al. "Part-of-speech tagger for Bodo language using deep learning approach." Natural Language Processing (2024): 1-15.

[8] Mitri, Aiom Minnette, et al. "Probing a pretrained RoBERTa on Khasi language for POS tagging." Natural Language Processing: 1-20.

[9] Tathagata Raha, Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2019. Development of POS tagger for English-Bengali Code-Mixed data. In Proceedings of the 16th International Conference on Natural Language Processing, pages 143–149, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

[10] Deka, Ridip Ranjan, Simanta Kalita, Kishore Kashyap, Manash P. Bhuyan, and Shikhar Kr Sarma. "A study of t'nt and crf based approach for pos tagging in assamese language." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 600-604. IEEE, 2020.

[11] Cing, Dim Lam, and Khin Mar Soe. "Improving accuracy of part-of-speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language." *International Journal of Electrical and Computer Engineering* 10, no. 2 (2020): 2023.

[12] Dalai, Tusarkanta, Tapas Kumar Mishra, and Pankaj K. Sa. "Part-of-speech tagging of Odia language using statistical and deep learning based approaches." *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, no. 6 (2023): 1-24.

[13] Appidi, Abhinav Reddy, Vamshi Krishna Srirangam, Darsi Suhas, and Manish Shrivastava. "Creation of corpus and analysis in code-mixed Kannada-English social media data for POS tagging." In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pp. 101-107. 2020.

[14] Shekhar, Shashi, Dilip Kumar Sharma, and M. M. Beg. "An effective bi-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and Roman Hindi." *Computación y Sistemas* 24, no. 4 (2020): 1415-1427.

[15] Bhattu, S. Nagesh, Satya Krishna Nunna, Durvasula VLN Somayajulu, and Binay Pradhan. "Improving code-mixed POS tagging using code-mixed embeddings." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, no. 4 (2020): 1-31.