Comparative Evaluation of Rule-Based, SVM, CRF and BiLSTM Models for Gujarati Part-of-Speech Tagging Using an Art and Culture Corpus

Dr. Pooja Bhatt[1], Dr. Pawan Whig[2],

[1] Postdoctoral Researcher; [2] Supervisor bhattpooja.393@gmail.com, pawan.whig@vips.edu

---

Abstract: Part-of-Speech (POS) tagging is a fundamental Natural Language Processing (NLP) task that assigns grammatical categories to words in a sentence. Gujarati, being a low-resource language, has limited benchmark datasets and standardized evaluation frameworks. This study presents a comparative analysis of Rule-Based, Support Vector Machine (SVM), Conditional Random Fields (CRF), and Bidirectional Long Short-Term Memory (BiLSTM) models for Gujarati POS tagging using an Art and Culture corpus containing approximately 20,000 manually annotated tokens based on BIS tag standards. A unified experimental framework is employed to ensure fair comparison across all models. The results show that BiLSTM achieved the highest accuracy of 96.7%, followed by CRF (94.2%), SVM (91.5%), and the Rule-Based approach (79.1%). The study demonstrates that deep learning models provide superior contextual understanding, while CRF remains an effective choice for low-resource Gujarati POS tagging.

Keywords: Gujarati NLP, POS Tagging, Art and Culture Corpus, CRF, BiLSTM, Low-Resource Languages

2. Literature Review

Part-of-Speech (POS) tagging has been extensively studied for high-resource languages; however, research on Gujarati remains relatively limited due to the scarcity of annotated corpora and linguistic resources. Over the past two decades, researchers have explored rule-based, statistical, machine learning, and deep learning approaches to improve tagging performance for Gujarati and other Indian languages.

2.1 Rule-Based Approaches

One of the earliest efforts in Gujarati POS tagging was undertaken by Patel and Gali (2008) in the paper *"Part-of-Speech Tagging for Gujarati Using Rule-Based Linguistic Techniques"*. The authors developed a handcrafted rule set based on Gujarati grammar, suffix patterns, and lexical resources. Although the approach was computationally inexpensive and interpretable, its performance was limited by linguistic ambiguity and the difficulty of maintaining extensive rule sets.

2.2 Machine Learning Approaches

Machine learning techniques have significantly improved POS tagging accuracy over rule-based methods**. Prajapati and Yajnik (2019)** demonstrated that Support Vector Machines (SVM) effectively use contextual features to improve tagging performance. Kumar et al. (2018) also reported that supervised machine learning models achieve robust results for morphologically rich Indian languages, although they require extensive feature engineering.

2.3 Conditional Random Field Based Approaches

Conditional Random Fields (CRFs) have become widely used for POS tagging because they effectively capture contextual dependencies in sequential text. **Patel et al. (2012)** demonstrated that CRF-based Gujarati POS tagging significantly outperformed traditional classifiers by utilizing contextual features. The foundational work of Lafferty, McCallum, and Pereira (2001) established CRFs as a powerful probabilistic framework for sequence labeling, making them a standard approach for POS tagging across multiple languages.

2.4 Deep Learning Approaches

Recent advances in deep learning have significantly improved POS tagging by reducing the need for handcrafted features**. Graves (2012)** demonstrated that Bidirectional Long Short-Term Memory (BiLSTM) networks effectively capture both preceding and succeeding contextual information, making them highly suitable for sequence labeling tasks. Studies on Indian and other low-resource languages have shown that BiLSTM models consistently outperform traditional machine learning techniques, particularly for morphologically rich languages such as Gujarati.

2.5 Research Gap

Despite significant progress in Gujarati POS tagging, several limitations remain:

1. Most studies evaluate only a single tagging approach rather than providing a comprehensive comparison.
2. Domain-specific corpora, particularly those related to Gujarati Art and Culture, have received limited attention.
3. Existing studies often employ different datasets and evaluation methodologies, making direct comparison difficult.
4. Few investigations compare linguistic, machine learning, and deep learning approaches under a unified experimental framework.
5. Limited research exists on the computational trade-offs between Rule-Based, SVM, CRF, and BiLSTM models for Gujarati POS tagging.

Reference	**Paper Title**	**Method**	**Reported Accuracy**
Patel & Gali (2008)	Part-of-Speech Tagging for Gujarati Using Rule-Based Linguistic Techniques	Rule-Based	75–80%
Patel et al. (2012)	Gujarati Part-of-Speech Tagging Using Conditional Random Fields	CRF	92–94%
Prajapati & Yajnik (2019)	Part-of-Speech Tagging for Gujarati Using Support Vector Machines	SVM	89–91%
Graves (2012)	Supervised Sequence Labelling with Recurrent Neural Networks	BiLSTM	95%+

Table 1. Literature Review

3. Proposed Methodology

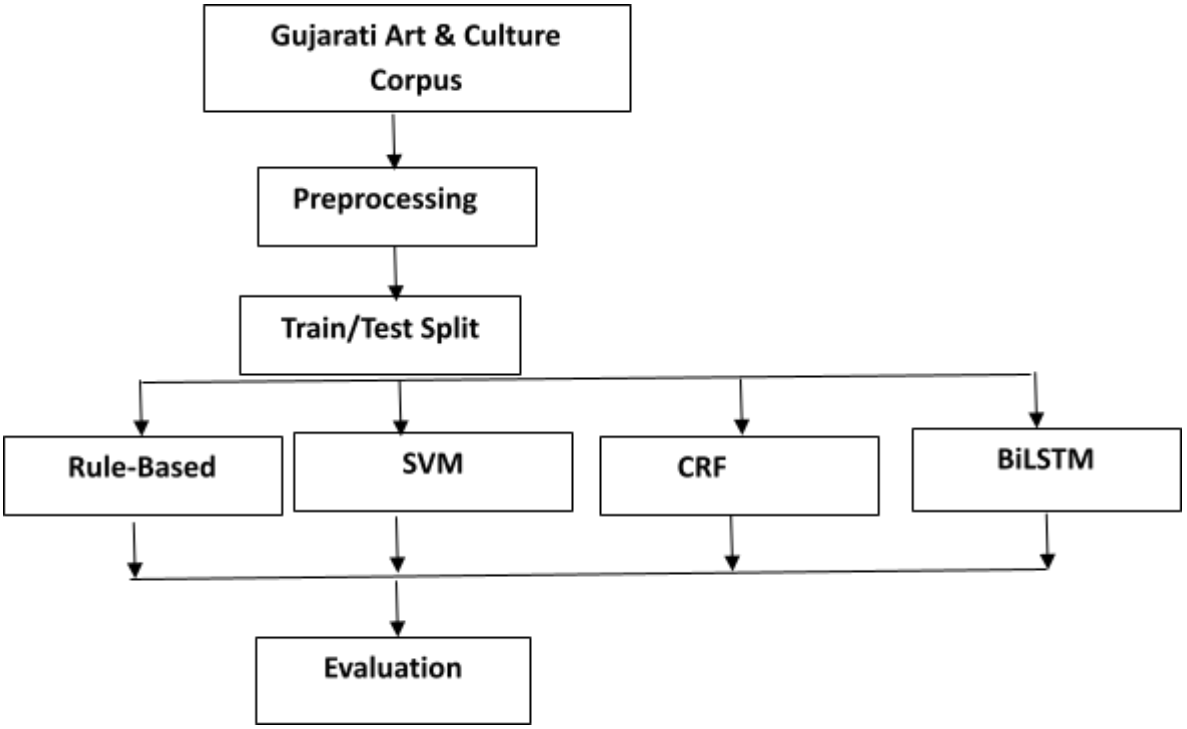


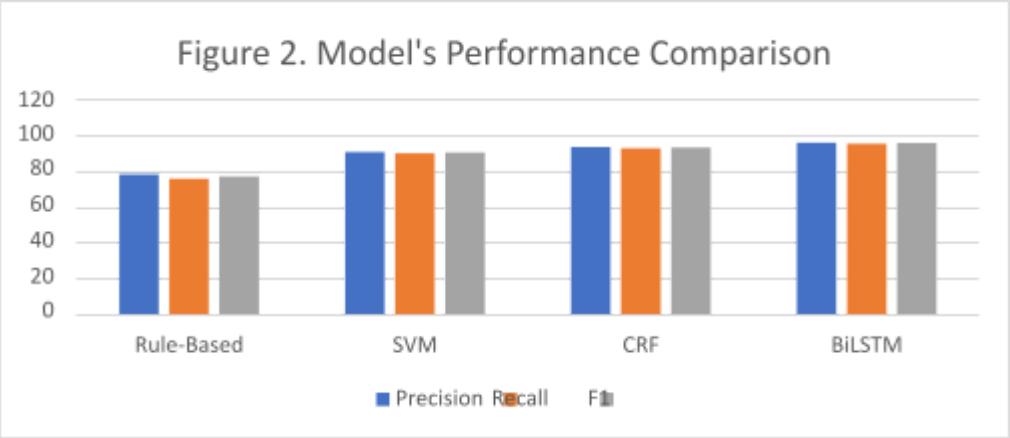
Figure 1. Proposed Methodology

The proposed methodology begins with the collection of a Gujarati Art & Culture corpus, followed by preprocessing steps such as tokenization, text normalization, and feature preparation. The processed dataset is then divided into training and testing sets to ensure a fair and consistent evaluation. Four POS tagging approaches—Rule-Based, Support Vector Machine (SVM), Conditional Random Fields (CRF), and BiLSTM—are trained and tested using the same dataset. Finally, the performance of all models is evaluated and compared using standard metrics such as Accuracy, Precision, Recall, and F1-score to identify the most effective approach for Gujarati POS tagging.

4. Results and Discussion

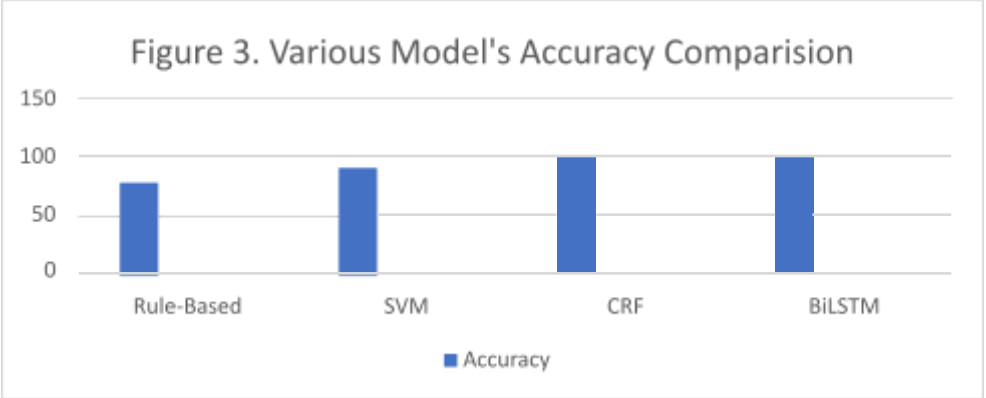
4.1 Performance Comparison

Figure 2 compares the precision, recall, and F1-score of the four POS tagging models evaluated on the Gujarati Art & Culture corpus. The results indicate that the BiLSTM model achieved the best overall performance, followed by CRF, while SVM demonstrated satisfactory performance and the Rule-Based approach recorded the lowest scores. This comparison highlights the effectiveness of deep learning models in capturing contextual information and improving POS tagging accuracy for Gujarati text.



4.2 Accuracy Comparison

Figure 3 presents the accuracy comparison of the four POS tagging models on the Gujarati Art & Culture corpus. The BiLSTM model achieved the highest accuracy (96.7%), followed by CRF (94.2%), SVM (91.5%), and the Rule-Based approach (79.1%), demonstrating that deep learning and sequence-based models outperform traditional rule-based techniques for Gujarati POS tagging.



Conclusion

This study presented a comparative evaluation of Rule-Based, SVM, CRF, and BiLSTM models for Gujarati POS tagging using a domain-specific Art and Culture corpus. Experimental results show that BiLSTM achieved the highest accuracy of 96.7%, outperforming CRF, SVM, and Rule-Based approaches. The findings demonstrate the effectiveness of neural architectures for Gujarati NLP while highlighting the continued relevance of CRF in low-resource settings. Future work will focus on larger corpora, character-level embeddings, and transformer-based architectures.

References

1. Bhatt, P., & Ganatra, A. (2021). POS-HOML: Part-of-speech tagging technique for Gujarati language using hybrid optimal and machine learning approaches. Journal of Ambient Intelligence and Humanized Computing, 12(12), 10489–10505.
2. Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Springer. https://doi.org/10.1007/978-3-642-24797-2
3. Gupta, V., & Lehal, G. S. (2010). A survey of part-of-speech tagging techniques for Indian languages. Journal of Emerging Technologies in Web Intelligence, 2(3), 169–173.

4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
5. Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd ed.). Pearson Education.
6. Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (ICML), 282–289.
7. Patel, C., et al. (2008). Gujarati POS tagging using conditional random fields., Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages(IJCNLP-08),117–122.
8. Prajapati, M., & Yajnik, A. (2019). POS tagging for Gujarati using machine learning approaches, Proceedings of the International Conference on Machine Learning and Data Science,102–108.