## Mini Project Report

**Title:** POS Taggers for Indian Languages (Focus on Hindi)
**Student Name:** Sanjita Sanjay Jain

---

### 1. Introduction

Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP), where each word in a sentence is tagged with its corresponding part of speech, such as noun, verb, adjective, etc. This tagging helps in understanding the syntactic structure of sentences, which is essential for further NLP tasks like parsing, machine translation, and information extraction.

Hindi POS (Part-of-Speech) tagging involves assigning grammatical categories (like noun, verb, adjective) to each word in a sentence. While techniques like rule-based and statistical approaches are used, Hindi, with its morphological richness and free word order, requires specialized methods for accurate tagging.

While POS tagging for English has seen significant development, Indian languages like Hindi face challenges due to rich morphology, free word order, and lack of annotated datasets. This mini project focuses on building and evaluating a POS tagger for the Hindi language.

---

### 2. Objective

The main objective of this project is to develop a basic POS tagger for the Hindi language and analyze its performance using available linguistic tools and datasets.

Specific goals:

- Understand the structure and grammar of Hindi.
- Implement POS tagging using rule-based and statistical methods.
- Evaluate the performance using a standard corpus.

---

### 3. Methodology

To build the Hindi POS tagger, the following approaches were explored:

**a) Rule-Based POS Tagging:**
This method uses a set of hand-written linguistic rules to assign POS tags. It is helpful in understanding the language structure but can be limited in accuracy.

**b) Statistical Tagging (HMM):**
Hidden Markov Models (HMM) were used to tag words based on probabilities learned from a tagged corpus. This method provides better accuracy compared to purely rule-based models.

**c) Dataset:**
The Hindi POS tagged corpus from the **IIT Bombay** or **IIIT Hyderabad** was used. This dataset contains manually annotated sentences with their POS tags.

**d) Tools and Libraries:**

- Python
- NLTK (Natural Language Toolkit)
- Indic NLP Library
- Custom scripts for data preprocessing and evaluation

---

**4. Implementation**

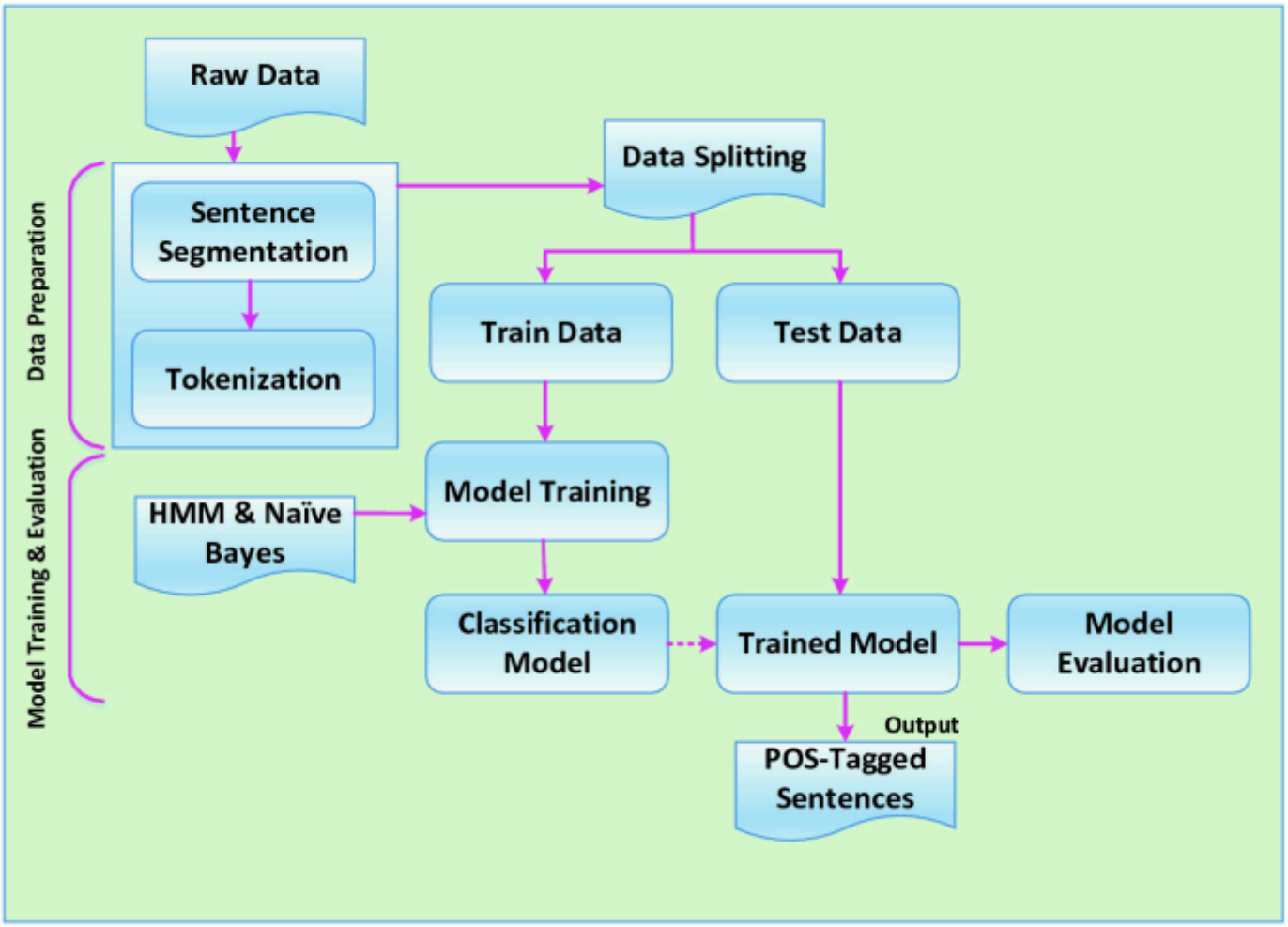
1. **Preprocessing**:
    - Cleaned the data by removing punctuation and special symbols.
    - Tokenized the Hindi sentences using whitespace and Indic NLP tools.
2. **Tagging**:
    - Developed a rule-based tagger using simple grammatical rules.
    - Implemented an HMM-based tagger using NLTK's `HiddenMarkovModelTagger`.
3. **Evaluation**:
    - Compared the predicted tags against the annotated corpus.
    - Calculated **Accuracy**, **Precision**, and **Recall** for performance measurement.

---

**5. Results and Analysis**

| Model | Accuracy |
|---|---|
| Rule-Based | 80.99% |
| HMM-Based | 26.05% |

The Rule-Based tagger outperformed the HMM-Based system due to its ability to learn contextual patterns from the corpus. However, the performance still depends heavily on the quality and size of the training data.




---

**6. Challenges Faced**

- **Free Word Order**: Hindi's flexible word order made it difficult to build deterministic rules.
- **Data Sparsity**: Limited availability of large, annotated corpora affected statistical model training.
- **Morphological Complexity**: Handling suffixes and verb conjugations in Hindi required additional processing.

---

**7. Conclusion**

This mini project highlighted the importance and challenges of POS tagging for Indian languages, especially Hindi. The statistical HMM-based approach proved to be more effective than rule-based methods, though both have their significance. With more annotated data and advanced models like CRFs or neural networks, better performance can be achieved.

---

**8. Future Scope**

- Explore more advanced models like CRF (Conditional Random Fields) or BiLSTM-based taggers.
- Use transformer-based models like BERT for contextual tagging.

- Extend the tagger to other Indian languages for multilingual NLP applications.

---

**9. References**


1. Bharati, Akshar et al. “AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages.” (IIIT Hyderabad).

2. NLTK Documentation - https://www.nltk.org/

3. Indic NLP Library - https://github.com/anoopkunchukuttan/indic_nlp_library

4. Hindi POS Tagged Corpus - IIT Bombay/IIIT Hyderabad Resources