Eduvest – Journal of Universal Studies
Volume 6 Number 4, April, 2026
p- ISSN 2775-3735- e-ISSN 2775-3727

Corpus Development and Pos Tagging Evaluation for Riau Malay Dialect Using Hidden Markov Model in A Low-Resource Setting

Rifky Akbar Vetian*, Koko Handoko, Andi Maslan, Alfannisa Annurrullah Fajrin
Universitas Putera Batam, Indonesia
Email: rifky.vetian@puterabatam.ac.id*, Koko.Handoko@puterabatam.ac.id, andimaslan@puterabatam.ac.id, alfannisa@puterabatam.ac.id

Keywords
Part-of-Speech Tagging; Hidden Markov Model; Natural Language Processing; Low-Resource Language; Bahasa Melayu Dialek Riau; Corpus Annotation.

ABSTRACT
Natural Language Processing (NLP) plays a crucial role in enabling machines to process and understand human language. One of the fundamental tasks in NLP is Part-of-Speech (POS) tagging, which serves as the foundation for various downstream applications such as parsing, information extraction, and machine translation. However, the development of POS tagging models for low-resource languages remains a significant challenge due to the limited availability of annotated corpora. This study aims to develop a POS-tagged corpus for Bahasa Melayu Dialek Riau (BMDR) and evaluate the performance of a Hidden Markov Model (HMM) as a baseline approach for POS tagging. The dataset consists of approximately 600 sentences with around 10,000 tokens, which were manually annotated and validated using Inter-Annotator Agreement. The annotated corpus was then divided into training and testing sets with a ratio of 80:20. Experimental results show that the HMM model achieved an accuracy of 86.8%, with precision, recall, and F1-score values of 85.9%, 85.3%, and 85.6%, respectively. The results indicate that HMM remains a competitive approach for POS tagging in low-resource language settings. Error analysis reveals that lexical ambiguity, Out-of-Vocabulary (OOV) words, and limited training data are the primary factors affecting model performance. This research contributes by providing the first annotated POS corpus for BMDR, evaluating the effectiveness of HMM in a low-resource context, and offering insights into linguistic challenges in regional languages. Future work may explore larger datasets and advanced deep learning models to improve tagging performance.

INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has significantly improved the processing of unstructured data, particularly textual data (Akbar Vetian et al., 2026). In this context, Natural Language Processing (NLP) plays a crucial role in enabling machines to understand, analyze, and interpret human language automatically (Jurafsky & Martin, 2000),(Porikli et al., 2018). NLP has been widely applied in various applications such as machine translation, sentiment analysis, and information retrieval systems, making it an essential component in data-driven digital transformation (Devlin et al., 2020),(Brown et al., 2020)

One of the fundamental tasks in NLP is Part-of-Speech (POS) tagging, which involves assigning grammatical categories to each token in a sentence based on its syntactic context. POS tagging serves as a foundational step for many downstream NLP tasks, including parsing, named entity recognition, and information extraction (Manning & Schutze,

1999),(Panchendrarajan & Amaresan, 2018). Errors in this stage can significantly degrade the overall performance of NLP systems.

Various approaches have been developed to address POS tagging, ranging from rule-based methods to machine learning and deep learning techniques. Modern models such as BiLSTM-CRF have demonstrated high performance across different languages(Huang et al., 2015),(Ma & Hovy, 2016). However, these approaches require large, annotated datasets and substantial computational resources. In contrast, classical statistical models such as Hidden Markov Model (HMM) remain relevant, particularly for low-resource languages, due to their computational efficiency and lower data requirements (Rabiner, 2002), (Wicaksono & Purwarianti, 2010)

A major challenge in POS tagging is lexical ambiguity, where a single word can have multiple possible tags depending on its context (Toutanova et al., 2003). This issue becomes more complex in languages or dialects that lack standardized linguistic resources. Bahasa Melayu Dialek Riau (BMDR), as a regional dialect closely related to Bahasa Indonesia, currently lacks sufficient digital linguistic resources (Sumoko et al., 2021),(Gil, 2002). The absence of a POS-tagged corpus for BMDR presents a significant barrier to developing NLP technologies that accurately reflect local language variations. Furthermore, the performance of POS tagging models heavily depends on the availability of large, annotated corpora, making it particularly challenging for low-resource languages (Joshi et al., 2020).

Indonesia, with more than 700 regional languages, faces significant challenges in developing NLP technologies for local languages. Although NLP research for Bahasa Indonesia has progressed considerably (Kurniawan & Aji, 2018), regional dialects remain underexplored, especially in terms of annotated corpus development. BMDR is an important dialect that serves as a lingua franca in its region and shares historical ties with Bahasa Indonesia. However, to date, there is no widely available POS-tagged corpus for BMDR (Sumoko et al., 2021).

Previous studies have shown that HMM remains an effective approach for POS tagging in low-resource settings. Research on the Madurese language demonstrated that HMM with the Viterbi algorithm achieved satisfactory accuracy (Dewi & Ubaidi, 2018), while studies on Malay Pontianak indicated that HMM performance is competitive compared to other models such as Conditional Random Fields (CRF) and Support Vector Machines (SVM) (Sumoko et al., 2021). Nevertheless, studies focusing on corpus development and POS tagging evaluation for BMDR are still very limited.

Based on these challenges, this study aims to develop a POS-tagged corpus for Bahasa Melayu Dialek Riau and evaluate the performance of the Hidden Markov Model (HMM) as a baseline approach for POS tagging. This research contributes by providing the first annotated POS corpus for BMDR, evaluating the effectiveness of HMM in a low-resource language context, and offering insights into linguistic characteristics and challenges of regional dialect processing.

METHOD

This study adopted a *design science research* approach, focusing on the development of an annotated corpus and the evaluation of a Part-of-Speech (POS) tagging model. The research

\

workflow consists of several stages, including data collection, preprocessing, corpus annotation, model construction, and performance evaluation.

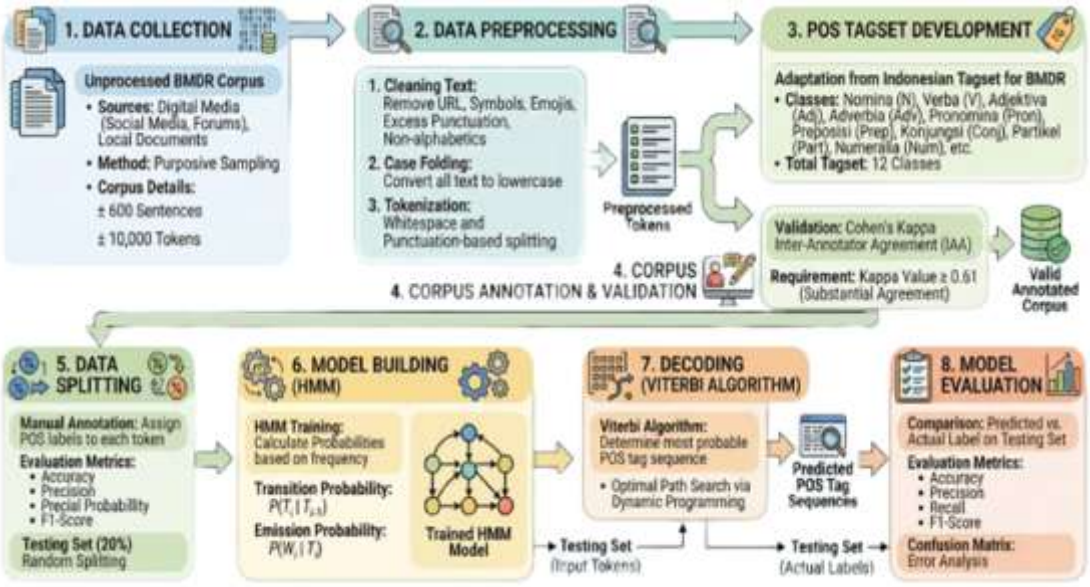

Figure 1. Research Methodology Flow POS Tagging Malay Language Riau Dialect

Figure 1 illustrates the overall research methodology, starting from data collection, preprocessing, corpus annotation, model training using Hidden Markov Model (HMM), and evaluation using standard performance metrics.

Dataset and Data Collection

The dataset used in this study consists of a corpus of Bahasa Melayu Dialek Riau (BMDR), collected from various digital sources such as social media, online forums, and local documents that reflect everyday language usage. The developed corpus includes:

Number of sentences : approximately 600 sentences

Number of tokens : approximately 10,000 tokens

A purposive sampling approach was applied to ensure the representation of diverse lexical variations and sentence structures within BMDR.

Preprocessing Data

Preprocessing was conducted to improve data quality prior to annotation and model training. The steps include:

1) Text Cleaning
   Removal of URLs, symbols, emojis, excessive punctuation, and non-alphabetic characters.
2) Case Folding
   Conversion of all text into lowercase to avoid token duplication.
3) Tokenization
   Segmentation of sentences into tokens using whitespace and punctuation-based rules.

POS Tag set Definition

The POS tag set used in this study is adapted from the Indonesian POS tag set, with adjustments to accommodate the linguistic characteristics of BMDR. The tag set includes:

- Noun (N)

- Verb (V)
- Adjective (Adj)
- Adverb (Adv)
- Pronoun (Pron)
- Preposition (Prep)
- Conjunction (Conj)
- Particle (Part)
- Numeral (Num)

In total, 12 POS categories are defined.

Corpus Annotation and Validation

The corpus was manually annotated by two annotators with expertise in Malay linguistic structures. Each token was assigned a POS label based on its syntactic role.

To ensure annotation consistency, Inter-Annotator Agreement (IAA) was measured using Cohen's Kappa coefficient. The interpretation of Kappa values is as follows:

- 0.61 - 0.80 → substantial agreement
- 0.80 → almost perfect agreement

The corpus is considered reliable when the Kappa value is ≥ 0.61.

Data Splitting

The annotated dataset was divided into:

- Training set : 80%
- Testing set : 20%

The split was performed randomly to minimize data bias and ensure fair evaluation.

Model Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is employed as the core approach for POS tagging. In this model, the sequence of POS tags is treated as hidden states, while the observed words represent the observable sequence.

Two main probabilistic components are defined:

- Transition Probability

  The probability of transitioning from one tag to another:

  $P(T_i|T_{i-1})$

- Emission Probability

  The probability of a word being generated from a specific tag:

  $P(W_i|T_i)$

where:

P = *Probability*

W = *Word*

T = *Tag*

i = Position Indexing

\

All probabilities are estimated based on frequency counts from the training dataset. Figure 2 illustrates the structure of the HMM used in this study.

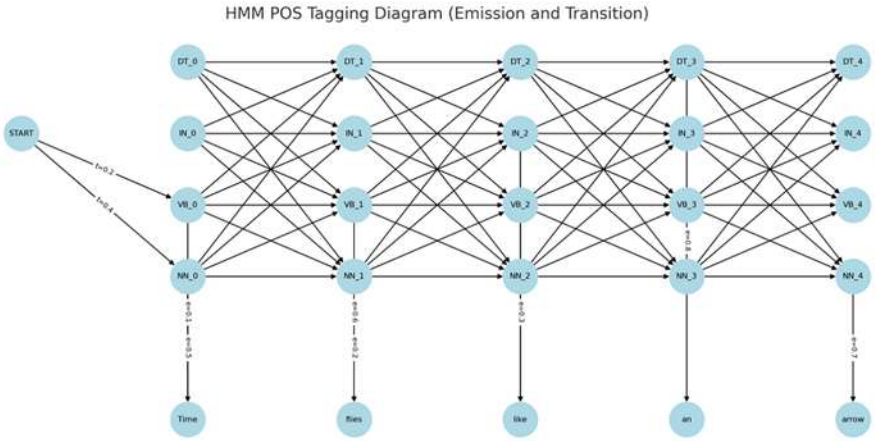

Figure 2. Hidden Markov Model Architecture in POS Tagging

The HMM model consists of two main components: hidden states (word classes) and observations (words). Transition probabilities indicate the relationship between tags, while emission probabilities indicate the relationship between tags and words.

Viterbi Algorithm

The Viterbi algorithm is used to determine the most probable sequence of POS tags for a given sentence. It employs a dynamic programming approach to efficiently compute the optimal path through the HMM state space.

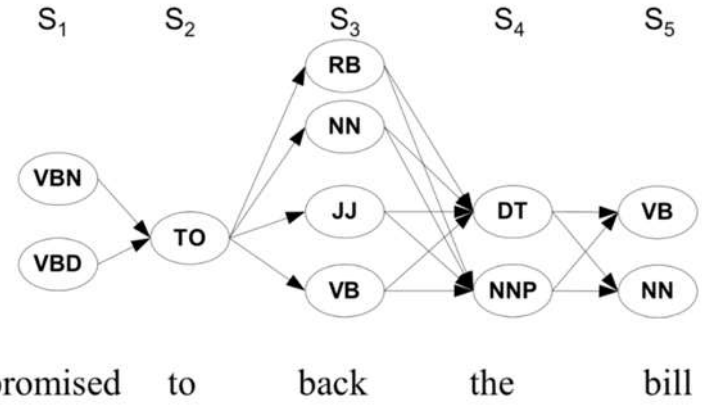

Figure 3. Illustration of the Viterbi Algorithm Process for Determining Tag Sequence

The Viterbi algorithm is used to determine the tag sequence with the highest probability through a dynamic programming process on a trellis structure.

Model Evaluation

The performance of the model is evaluated by comparing predicted POS tags with the ground truth labels in the testing dataset. The evaluation metrics used include:

- Accuracy
- Precision

- Recall
- F1-Score

In addition, a confusion matrix is utilized to analyze the distribution of classification errors across different POS categories.

RESULT AND DISCUSSION

Experimental Results

The proposed Hidden Markov Model (HMM) was evaluated using a testing dataset comprising 20% of the total corpus. The evaluation was conducted using standard performance metrics, including accuracy, precision, recall, and F1-score, to assess the effectiveness of the model in performing POS tagging on Bahasa Melayu Dialek Riau (BMDR). The evaluation results are presented in Table 1.

Table 1. Performance Evaluation of the HMM Model

Metrics	Value
Accuracy	86.8%
Precision	85.9%
Recall	85.3%
F1-Score	85.6%

Source: Processed research data (2026)

The results indicate that the HMM model achieves a relatively strong performance, with an accuracy of 86.8%, demonstrating that the majority of tokens were correctly classified. The balanced values of precision, recall, and F1-score further suggest that the model performs consistently across different POS categories without significant bias toward specific classes.

Confusion Matrix Analysis

To further analyze classification performance, a confusion matrix is presented in Figure 4.

The confusion matrix shows that most correct predictions are concentrated along the diagonal, indicating that the model is generally effective in distinguishing between POS categories. However, several misclassification patterns are observed, particularly among linguistically similar classes.

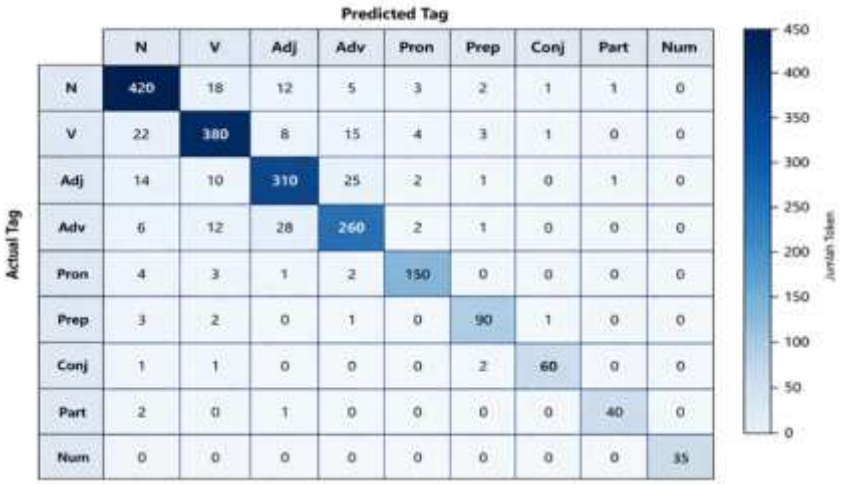

Figure 4. Confusion Matrix

\

The most prominent errors occur between:

- Noun (N) and Verb (V)
- Adjective (Adj) and Adverb (Adv)

These misclassifications can be attributed to the functional similarity and contextual overlap between these categories in BMDR, which pose challenges for probabilistic models such as HMM.

Error Analysis

The observed classification errors can be explained by several key factors:

1. Lexical Ambiguity
   Certain words in BMDR can belong to multiple POS categories depending on context. Since HMM relies on statistical probabilities, it has limitations in capturing complex contextual dependencies.
2. Out-of-Vocabulary (OOV) Words
   Words that do not appear in the training data lead to zero emission probabilities, reducing the model's ability to generate accurate predictions.
3. Limited Training Data
   The relatively small corpus size (approximately 10,000 tokens) limits the representativeness of probability distributions, affecting model generalization.

Comparison with Previous Studies

The performance of the proposed model is comparable to previous studies on low-resource languages.

- Madurese Language (HMM): ~86.03%
- Malay Pontianak (HMM): ~87%
- **BMDR (This Study): 86.8%**

These results indicate that HMM provides stable and competitive performance across different regional languages, even under limited data conditions.

The findings of this study confirm that classical statistical approaches such as HMM remain effective for POS tagging in low-resource language settings. While deep learning models such as BiLSTM-CRF generally achieve higher accuracy, their dependency on large, annotated datasets limits their applicability in regional language contexts.

Furthermore, the error analysis highlights that linguistic characteristics of BMDR such as lexical ambiguity and overlapping syntactic functions play a significant role in influencing model performance. This suggests that improving POS tagging performance for BMDR requires not only larger datasets but also more context-aware modelling approaches.

Future improvements may include:

1) expanding the size and diversity of the corpus
2) applying smoothing techniques to handle sparse data
3) integrating hybrid or deep learning-based models

CONCLUSION

This study successfully developed an annotated corpus for Bahasa Melayu Dialek Riau (BMDR) and implemented a Hidden Markov Model (HMM) for Part-of-Speech (POS) tagging

in a low-resource language setting. The constructed corpus consists of approximately 600 sentences with around 10,000 tokens, which were manually annotated and validated using the Inter-Annotator Agreement (IAA) method. The experimental results demonstrate that the proposed HMM model achieved an accuracy of 86.8%, with balanced precision, recall, and F1-score values. These findings indicate that the linguistic structure of BMDR exhibits sufficient regularity to be effectively modelled using probabilistic approaches. Furthermore, the results confirm that classical statistical methods remain competitive for POS tagging tasks in low-resource language contexts. However, several limitations were identified. The presence of lexical ambiguity, functional overlap between POS categories, and the limited size of the training dataset restrict the model's ability to capture contextual dependencies effectively. Additionally, the occurrence of Out-of-Vocabulary (OOV) words contributes to performance degradation. The main contributions of this study are threefold, the development of the first POS-tagged annotated corpus for BMDR. The evaluation of HMM performance in a regional low-resource language context. And, an error analysis that provides insights into the linguistic characteristics and challenges of BMDR. For future work, it is recommended to expand the size and diversity of the corpus and to explore advanced approaches such as BiLSTM-CRF or transformer-based models to improve tagging performance. Additionally, the implementation of smoothing techniques and more robust OOV handling strategies is expected to further enhance model effectiveness.

REFERENCE

Akbar Vetian, R., Fajrin, A. A., & Maslan, A. (2026). Penerapan Model BERT dalam Klasifikasi Otomatis Laporan Pemeliharaan Industri Menggunakan Natural Language Processing. In *Jurnal Kecerdasan Buatan dan Data Science Industri: I* (Number 1).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. http://arxiv.org/abs/2005.14165

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2020). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. Retrieved https://github.com/tensorflow/tensor2tensor

Dewi, N. P., & Ubaidi, U. (2018). Lexical Rule and Lexicon Effect for Part of Speech Tagging Bahasa Madura. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *18*(1), 65–72. https://doi.org/10.30812/matrik.v18i1.332

Gil, D. (2002). Riau Indonesian Sama: a Unified Analysis. *NUSA*, *50*.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv Preprint ArXiv:1508.01991*. http://arxiv.org/abs/1508.01991

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual*

\

Meeting of the Association for Computational Linguistics, 6282–6293. https://microsoft.github.io/linguisticdiversity

Jurafsky, Dan., & Martin, J. H. . (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall.

Kurniawan, K., & Aji, A. F. (2018). Toward a Standardized and More Accurate Indonesian Part-of-Speech Tagging. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 303–307. https://doi.org/10.1109/IALP.2018.8629236

Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *1*, 1064–1074.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

Panchendrarajan, R., & Amaresan, A. (2018). Bidirectional LSTM-CRF for Named Entity Recognition. *32nd Pacific Asia Conference on Language, Information and Computation*.

Porikli, F., Shan, S., Snoek, C., Sukthankar, R., & Wang, X. (2018). Deep Learning for Visual Understanding: Part 2 [From the Guest Editors]. In *IEEE Signal Processing Magazine* (Vol. 35, Number 1, pp. 17–19). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MSP.2017.2766286

Rabiner, L. R. (2002). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *2*(77), 257–286.

Sumoko, A., Negara, A. B. P., & Pratiwi, H. S. (2021). Perbandingan Tipe Metode PoS Tagger Terhadap Nilai Akurasi Untuk Bahasa Melayu Pontianak. *Jurnal Sistem Dan Teknologi Informasi (Justin)*, *9*(3), 342. https://doi.org/10.26418/justin.v9i3.44116

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL*, 252–259.

Wicaksono, A. F., & Purwarianti, A. (2010). HMM Based Part-of-Speech Tagger for Bahasa Indonesia. *On Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop*. http://students.itb.ac.id/home/alfan_fw@students.itb.ac.id/IPOSTAgger