Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

Constructing a Part-of-Speech Tagging based on Lexicon and Rule-based for Sundanese Corpus

Ade Sutedi[*1], Ayu Latifah[2], Novan Rodiansyah[3], Yayat Sudaryat[4]

[1,2,3]Institut Teknologi Garut, Indonesia
[4]Universitas Pendidikan Indonesia, Indonesia

Email: [1]adesutedi@itg.ac.id

Received : Sep 30, 2025; Revised : Jan 19, 2026; Accepted : Feb 10, 2026; Published : Jun 15, 2026

Abstract

Part-of-Speech (POS) Tagging is the process of annotating word classes (nouns, verbs, adjectives, etc.) in a sentence, which is used as a basis for natural language processing and artificial intelligence. In this study, a corpus of word classes and word class annotating rules for the Sundanese language, which has limited resources, was developed. The experiments were conducted on an annotated corpus consisting of 104,696 tokens collected from Sundanese dictionaries, Sundanese Literature (Carita Pondok, Guguritan, Mantra, Pupujian, Sisindiran, Sajak, and Wawacan), Babasan and Paribasa, and social media X (Twitter). The annotation process is carried out in several stages that combine manual annotation based on cross-lingual transfer from Indonesian POS to Sundanese POS, then adjusted based on the word class rules in Sundanese. The results of this study are a POS annotation corpus containing Sundanese word-tag pairs and a basic rule-based model compared to the HMM and CRF models. The rule-based model achieves an F1-score of 0.867, the CRF model achieves an F1-score of 0.889, while the HMM model attains the highest score with an F1-score of 1.000. Analysis of POS distributions reveals that nouns (KB) consistently dominate across all models, reflecting the noun-rich nature of Sundanese literary texts. It also highlights the challenges of handling unknown words and the need for richer annotated resources, which are related to tag interoperability with Universal POS standards. This research contributes to the development of NLP resources for low-resource languages and provides a methodological foundation for future Sundanese NLP applications.

Keywords : *Annotation, Corpus, Part-Of-Speech, Sundanese.*

1. INTRODUCTION

Part-of-speech (POS) tagging is one of the basic stages in Natural Language Processing (NLP) to label grammatical categories of words in a text, such as nouns, verbs, adjectives, and so on. POS tagging involves labelling each word in a text with its corresponding grammatical category. Given an input sequence of tokenized words $x_1, x_2, ..., x_n$ and a predefined tag set, the task produces an output sequence $y_1, y_2, ..., y_n$ where each tag y_i corresponding exactly to one input is directly aligned with the word x_i [1]. POS tagging plays an important role because it provides basic syntactic information that can be utilized in various NLP applications, including keyword extraction [2], dependency parsing [3], name entity recognition [4], machine translation [5], text generation [6], and question answering [7] .

In Indonesia, POS tagging has begun to be developed for Indonesian [8], [9], [10] as well as regional languages such as Javanese [11], [12], Madurese [13], [14], Malay [15], and other regions. In research related to part-of-speech (POS) tagging, methods have evolved from early rule-based approaches [13], [16] to probabilistic models [11], [12], [17], [18] and deep learning techniques [16], [19], [20]. Early work in POS tagging primarily relied on rule-based systems, such as the Brill tagger [13], which applies transformation-based learning to iteratively refine an initial tagging output using manually designed rules. Subsequently, probabilistic generative models became prominent, particularly

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

the Hidden Markov Model (HMM) [10], [11], [12], which models the joint probability of word and tag sequences. To efficiently decode the most probable tag sequence in HMMs, the Viterbi algorithm [14], based on dynamic programming, is commonly employed. As research progressed, discriminative sequence labelling models gained attention, notably Conditional Random Fields (CRF) [9], [17], [21], [22], which directly model the conditional probability of label sequences given observations and better capture dependencies between adjacent tags. More recently, deep learning approaches have become dominant in POS tagging. Neural architectures such as Convolutional Neural Networks (CNNs) [9] are used to capture local contextual and morphological patterns, while Bidirectional Long Short-Term Memory (BiLSTM) networks [9], [21], [22] effectively model long-range contextual dependencies in both forward and backward directions. In parallel, domain adaptation techniques [8] were introduced to improve model robustness across corpora with differing linguistic characteristics. Further advancements include contextualized word representation models such as Embeddings from Language Models (ELMo) [9], [23], which generate dynamic embeddings that adapt word meaning based on context and significantly enhance tagging performance.

However, regarding Sundanese in particular, POS tagging research is still limited and without a standard and adequately annotated corpus, resulting in society gaining little or no benefit from recent advances in natural language understanding [24]. In addition, regional languages often have complex morphology, including affixation, reduplication, and various word forms [25], thus making the process of automatic word class classification difficult. Dialect variations [13] and the lack of spelling standardization [26] also add complexity to tagger development. This is important because Sundanese is one of the regional languages with the largest number of speakers in Indonesia [27], so there is a high urgency for the availability of a POS tagging corpus. Therefore, this study aims to develop a corpus of word classes (POS tagging) in Sundanese, so that it can become a useful linguistic resource for further research. We apply and compare the POS tagging algorithm, with our Rule-based method as the base model, and compare the performance with the HMM model and the CRF model in tagging word classes in Sundanese texts.

2. METHOD

This research uses an experimental approach that begins with corpus collection, data annotation, modeling the post tag rules for Sundanese language, then comparing it with previous methods, and evaluating the POS tag results model. Figure 1 below shows the steps taken in the research in constructing POS tagging annotations for Sundanese language.

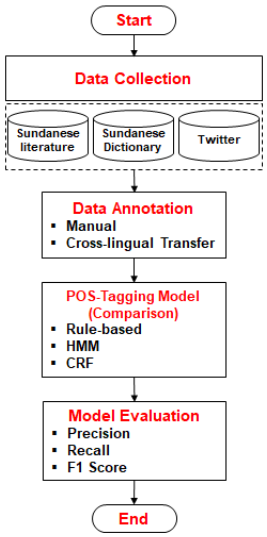

Figure 1. Research flow diagram

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

2.1. Data Collection

Data was collected from several sources, namely Sundanese dictionaries, Sundanese literary texts, and social media with total more than 100000 token. Data collection was carried out manually by several students and also crawling data for twitter (X) social media then writing a list of words into a file in txt format.

2.1.1. Sundanese Dictionary

In this study, one of the data used comes from the Sundanese dictionary written by R. A. Danadibrata [28]. The tokens used are general words that are used without including words in abbreviated form.

2.1.2. Sundanese Literatures

Furthermore, the dataset is taken from several types of Sundanese literature contained in textbooks [29], [30] including:

1. *Babasan*: a language phrase whose meaning and origin are fixed. The phrase *parondok* is a short phrase, usually consisting of only two words and containing the meaning of a proverb and describing human behavior.
2. *Paribasa*: a language phrase that is longer than an expression, generally containing a deeper meaning, some containing a call to action and some containing a warning.
3. *Carpon* (*Carita Pondok*): a short story that represents a unified idea. In its brevity and conciseness, a carpon is complete, rounded, and concise.
4. *Mantra*: a form of free verse containing supernatural powers, its use is not arbitrary. Mantras are usually recited by heart. The purpose is to use supernatural powers to achieve a goal.
5. *Wawacan*: a story in *dangding* form, written in *pupuh* poetry. The text of the *wawacan* is narrative, generally long, with frequent changes in *pupuh*, usually accompanying changing episodes.
6. *Guguritan*: a term used to refer to one or several stanzas of a poetic form that is usually sung, usually not long. This form of poetry is called *pupuh*, which consists of 17 types, namely *Kinanti, Asmarandana, Sinom, Dangdanggula, Pucung, Maskumambang, Magatru, Mijil, Wirangrong, Pangkur, Durma, Lambang, Gambuh, Balakbak, Ladrang, Jurudemung*, and *Gurisa*, each with its own rules, which essentially revolve around the provisions of (a) the number of lines in one stanza or pada, (b) the number of syllables in each line or *padalisan*, and (c) the vowel sound at the end of each line.
7. *Sisindiran*: Similar to *pantun* (Malay/Indonesian). Consists of two parts: *sampiran* (shell) and content (*eusi*). Usually four lines (or even), also known as *susualan* (riddles) and *bangbalikan* (content behind the *sampiran*).
8. *Pupujian*: Poetry containing praise, prayers, advice, and Islamic teachings. Living in a pesantren (Islamic boarding school) and religious study environment.
9. Poetry: A branch of literature that uses words to convey illusions, imagination, and ideas, much like a painting with lines and colors.

2.1.3. Social Media

Dataset from the social media Twitter (X) is also used as a reflection of everyday language use directly today. The data taken through the crawling process as many as that contain various forms of language expressions, ranging from formal to non-formal sentences, including abbreviations, emoticons, and mixed languages that cause ambiguity, so it becomes a unique challenge in the POS Tagging process.

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

2.2. Part-of-Speech (POS)

Part-of-Speech (POS) or syntactic category is the process of classifying word classes [8], [9] (word colors) in the form of grammatical units based on their form, properties, and behavior in a construction [25]. The separation of word colors is determined based on their form, properties, function, and behavior in the flow of sentences or syntactic constructions which are divided into two colors, namely the head word and the subject word. The head word (main) is the color of words that have lexical meaning, are sensitive to nature, culture, and place, and can generally be changed in shape. The subject word (means, particles) is the color of words that generally serve as sentence tools, usually do not have lexical meaning, but grammatical meaning, and are difficult to change in shape. The color of the head word and the subject word has sub-class. The sub-class may not be the same at all, they have their own characteristics, but are still a group. However, the sub-class of words are very similar to the main word.

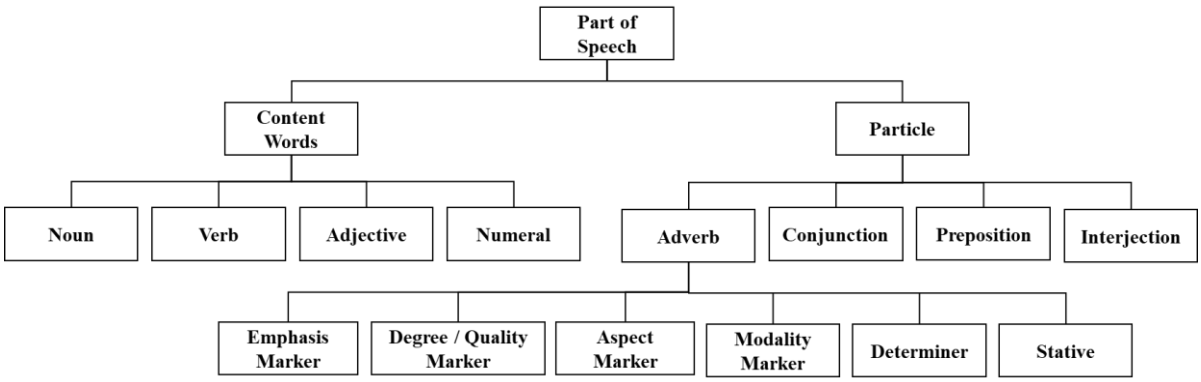

Figure 2. Sundanese word classification

In general, the class of words in Sundanese can be divided in several section which depicted in Figure 2. Based on this word class division, the POS tag pattern in Sundanese is similar to both Indonesian and English. Therefore, to determine the POS tag for a Sundanese word token, we can adapt existing POS tag patterns and adapt them to other terms that apply according to Sundanese word class rules.

2.3. Data Annotation

In this research, data annotation, or word labeling is a fundamental step, particularly in POS tagging. These annotations enable systems to understand the syntactic and semantic structure of a language, including low-resource languages like Sundanese. Various approaches can be used, ranging from manual methods to statistical models and machine learning.

2.3.1. Manual Annotation

In this research, manual methods involve human annotators directly labeling word classes within a corpus. The advantage of this method is its high level of accuracy, as it takes into account linguistic context and complex grammatical rules. However, its disadvantages include the high cost and time requirements, making it difficult to implement on a large corpus scale [1]. In the Table 1 presents a list of POS tag patterns for English, Indonesian, and Sundanese.

2.3.2. Cross-lingual Transfer

Cross-lingual transfer approaches utilize POS models trained on a source language (e.g., Indonesian or English), then transferred to a target language in this case is Sundanese. Studies have shown that cross-lingual POS transfer can improve performance in resource-constrained languages [5],

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

[7], [8]. We adopted Sundanese monolingual corpus [33] to enrich words in the lexicon dataset that will be used for the Sundanese language POS tagging corpus.

Table 1. Word class (POS Tag)

English [31]		**Indonesian** [9]		**Sundanese** [32]
Traditional POS	**Universal POS**	**Traditional POS**	**ID POS**	**Sun POS**
Noun	NOUN	Coordinating	CC	NOUN = KB (*Kecap Barang*)
	PROPN	conjunction	CD	PROPN = KN (*Kecap Nami*)
Verb	VERB	Cardinal number	OD	VERB = KP *(Kecap Pagawean*)
	AUX	Ordinal number	DT	AUX = KBT (*Kecap Bantu*)
Adjective	ADJ	Determiner / article	FW	ADJ = KS (*Kecap Sipat*)
	DET	Foreign word	IN	DET = PNJ (*Panunjuk*)
	NUM	Preposition	JJ	NUM = WIL (*Wilangan*)
Adverb	ADV	Adjective	MD	ADV = KT (*Kecap Katambah*)
Pronoun	PRON	Modal and auxiliary	NEG	ADP = PA (*Panganteur*)
Preposition	ADP	verb	NN	CCONJ = KPN (*Kecap*
Conjunction	CCONJ	Negation	NNP	*Pangantet*)
	SCONJ	Noun	NND	SCONJ = KPB (*Kecap*
Interjection	INTJ	Proper noun		*Pangabanding*)
	PART	Classifier, partitive,	PR	PART = PL (*Partikel*)
	X	and measurement	PRP	PRON = KG (*Kecap Ganti*)
	SYM	noun	RB	X = X (Others)
	PUNCT	Demonstrative	RP	PUNCT = TB (Tanda Baca)
		pronoun	SC	
		Personal pronoun	SYM	
		Adverb	UH	
		Particle	VB	
		Subordinating	WH	
		conjunction	X	
		Symbol	Z	
		Interjection		
		Verb		
		Question word		
		Unknown		
		Punctuation		

2.3.3. Rule Based

Rule-based methods use explicit linguistic rules, such as affix patterns or word order, to determine word classes. In local languages rich in affixes such as Sundanese, this method is quite effective in recognizing morphological patterns. For example, research on Madurese using rule-based Brill Tagger has achieved quite good results [13]. However, its main weakness is the difficulty in covering everyday language variations, including dialects and slang that often appear on social media [2], [3]. To strengthen the novelty of the research, we developed our own rule-based technique by referring to the rules of word classes in Sundanese [32].

2.3.4. Hidden Markov Model

Hidden Markov Model (HMM) is a generative statistical approach widely used in POS tagging research, especially in regional languages with limited resources. Mursyit et al. [11] applied HMM to Javanese word class labeling and showed that modeling transition probabilities between tags and word emission probabilities to tags can effectively capture linguistic sequence patterns. A similar approach was also carried out by Pratama et al. [12], where HMM provided stable performance despite the limited

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

amount of training data and the high morphological variation of the language. As a comparison, Cahyani and Mustikaningtyas [10] used Maximum Entropy Markov Model (MEMM) in Indonesian, which, although discriminative, remains within the Markov model framework and emphasizes the importance of sequential dependency modeling in POS tagging tasks. Thus, HMM can be positioned as a strong and relevant baseline model for POS tagging, especially in regional languages, as well as a starting point for the development of more complex models.

2.3.5. Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) offer improvements over HMMs because they are able to consider more complex contextual features and do not assume strict independence [1]. CRFs have been shown to excel in POS tagging of local languages. For example, research on Javanese has shown that CRFs can improve accuracy compared to HMM methods [17]. Furthermore, the combination of CRFs with modern word representations has been shown to strengthen POS tagging performance in Indonesian [9], and can be adapted to Sundanese with promising results.

3. RESULT

This study shows that word class labeling (POS tagging) for Sundanese requires adjustments to the morphological structure and vocabulary typical of the regional language. By using adapted POS patterns, such as NOUN = KB (*Kecap Barang*), PROPN = KN (*Kecap Nami*), and VERB = KP (*Kecap Pagawean*), the system can recognize word categories more precisely according to the Sundanese linguistic context. Other important categories such as AUX = KBT (*Kecap Bantu*), ADJ = KS (*Kecap Sipat*), and PRON = KG (*Kecap Ganti*) help clarify syntactic roles in sentences. The distribution of POS depicted in Figure 3.

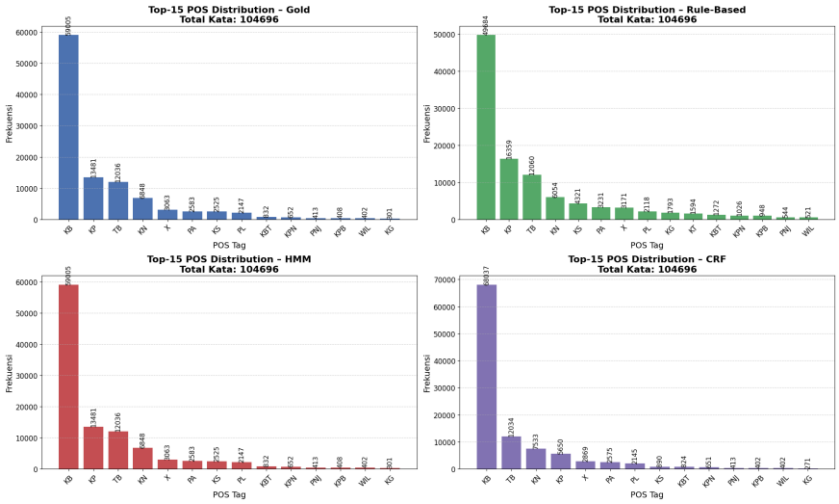

Figure 3. Sundanese POS tag distribution result

The POS tagging technique in this study uses an approach that combines lexicon-based methods by matching each word with a dictionary containing word pairs and POS labels. If a word is found in the lexicon, a rule-based approach is used that utilizes morphological rules in Sundanese, such as the suffixes "-keun" and "-an" for verbs, and "-na" and "-eun" for nouns. Next, the HMM approach is used as a sequence-based statistical method. In addition, the CRF approach is also used to handle words with a data division of 80% for training and 20% for testing, by utilizing contextual features such as word forms. The lexicon and rule-based methods, HMM, and CRF are then compared to determine the extent to which the results of the three approaches compare. Thus, the results of this study can be used as a baseline for further research.

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

In this research POS tagging performance evaluation is carried out by comparing the predicted labels with the ground truth labels on the test data using Precision (Eq. 1), Recall (Eq. 2), and F1-Score (Eq. 3) metrics. Precision is used to measure how accurate a model is in making predictions. This metric shows how many of the predicted positive results are actually correct, which is determined by comparing the number of true positives (TP) to the total of true positives (TP) and false positives (FP). A high precision value indicates that the model rarely makes incorrect predictions. Recall is used to measure the ability of the model to identify all data that truly belongs to a class. This metric compares the number of true positives (TP) to the total number of actual positive data, which includes true positives (TP) and false negatives (FN). A high recall value indicates that the model is able to detect most relevant data and misses only a small amount. The F1-Score is a combined metric that balances precision and recall by considering true positives (TP), false positives (FP), and false negatives (FN) together.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (3)$$

Based on the evaluation results with rounding to three decimal places, the Rule-Based model showed the best performance compared to other models, with a precision value of 0.900, a recall of 0.850, and an F1-score of 0.867. These results indicate that the rule-based approach is able to provide a good balance between prediction accuracy and completeness. Meanwhile, the HMM model obtained a precision value of 1.000, a recall of 1.000, and an F1-score of 1.000, which indicates a decrease in performance, especially in the recall aspect. The CRF model has relatively similar performance to the HMM, with a precision of 0.905, a recall of 0.905, and an F1-score of 0.889, but slightly lower in the F1-score. Overall, these results indicate that in this evaluation scenario, the HMM approach is more effective than statistical and sequential learning methods such as Rule-Based and CRF.

Table 2. Evaluation score

Model	Precision	Recall	F1
Rule-Based	0.900	0.850	0.867
HMM	1.000	1.000	1.000
CRF	0.905	0.905	0.889

4. DISCUSSIONS

The Sundanese POS Tag pattern has several advantages that make it relevant for regional language-based natural language processing. First, this system is contextual and local, as it is tailored to the morphological structure and vocabulary of Sundanese, making it easier for native speakers to understand. Second, this pattern can reduce ambiguity, particularly in distinguishing word functions that are difficult to accommodate with the Universal POS Tag, such as the use of typical Sundanese particles like -mah or -téa. Third, this pattern significantly supports the development of local NLP applications, such as machine translation, sentiment analysis, and Sundanese-based chatbots. Finally, this pattern also plays a crucial role in language preservation, as it brings traditional Sundanese terms into the computational realm and strengthens its linguistic identity.

Based on the POS distribution in Figure 3, it can be seen that all POS tagging approaches process the number of tokens, 104,696 words which is extracted from several Sundanese literatures. The POS KB (*Kecap Barang*) or NOUN in universal POS appears as the most dominant class. In the HMM results, KB frequencies are relatively balanced, while in the Rule-Based model, KB frequencies tend to

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

be lower and the CRF shows a significantly higher KB dominance than the other models. The POS KP (*Kecap Pagawean*) or VERB distribution shows a fairly consistent pattern in HMM. In the Rule-Based approach, the frequency of KP remains relatively high. Meanwhile, the CRF model shows a significant decrease in the KP class, indicating that the features used in the CRF model are not yet robust enough to capture the characteristics of Sundanese VERB, particularly affixation patterns such as prefixes and suffixes. The POS TB (*Tanda Baca*) or Punctuation Marks in both Rule-Based and HMM models maintaining distributions than the CRF. These findings emphasize the importance of selecting an appropriate approach and feature design in developing POS taggers for resource-constrained languages like Sundanese.

However, the Sundanese POS Tag pattern also has several drawbacks. First, this system lacks universality because it is not directly compatible with international standards, requiring additional mapping for cross-language integration. Second, its use is still limited to Sundanese, making it inflexible for application to other languages. Third, the availability of POS annotated datasets in Sundanese is still limited, which makes it difficult to develop and train statistical and deep learning models. Fourth, even though ambiguity is reduced, there are still cases of words in Sundanese that are multi-functional, such as the word alus which can function as an adjective or adverb, so that it requires additional disambiguation strategies.

5. CONCLUSION

This study concludes that POS tagging in Sundanese requires linguistic adaptations that explicitly reflect the morphological structure and vocabulary of the language to achieve reliable performance. The use of Sundanese-specific POS labels such as KB (*Kecap Barang*), KP (*Kecap Pagawean*), and KN (*Kecap Nami*), and etc. proved to be able to represent syntactic categories more contextually. The integration of lexicon and rule-based approaches yielded stable and balanced results, and enabled comprehensive comparisons with statistical approaches such as HMM and CRF in a resource-limited regional language context. Analysis of the POS distribution of 104,696 tokens showed that KB or NOUN consistently dominated across all models, reflecting the nominal-rich characteristics of Sundanese literary texts. From an evaluation perspective, each approach demonstrates distinct advantages, the Rule-Based model achieves a balanced performance with F1-score of 0.867, indicating its effectiveness in combining accuracy and coverage, the CRF model shows competitive performance with an F1-score of 0.889, and the HMM model achieves highest scores with F1-score 1.000. Although the results demonstrate strong potential, this study still has several limitations. First, the availability of Sundanese annotated data is still limited, thus limiting the model's generalizability across different domains. Second, the use of Sundanese-specific POS tags reduces cross-language interoperability and compatibility with the universal POS standard and needs to be further developed to mapping between Sundanese POS tags and Universal POS Tags. Third, there is still ambiguity in multifunctional words that can act as nouns, verbs, or other categories depending on context, which cannot be fully addressed by the rules or statistical features used. Future research directions include expanding the annotated corpus with a wider variety of domains, and exploring hybrid approaches that combine linguistic rules with deep learning models such as BiLSTM or Transformer.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Prof. Dr. Yayat Sudaryat, M. Hum. for valuable guidance and constructive feedback throughout this research. We also thank Ministry of Higher Education, Science, and Technology Republic of Indonesia through research contracts number 125/C3/DT.05.00/PL/2025, Higher Education Service Institution (LLDIKTI IV) through research contracts number 8004/LL4/PG/2025, and Institut Teknologi Garut through research contracts number

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

477/ITG/A.11/B/VI/2025 for providing the funding resource and support necessary to complete this study.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed. 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[2] E. Altuncu, J. R. C. Nurse, Y. Xu, J. Guo, and S. Li, "Improving Performance of Automatic Keyword Extraction (AKE) Methods Using PoS Tagging and Enhanced Semantic-Awareness," *Information*, vol. 16, no. 7, p. 601, Jul. 2025, doi: 10.3390/info16070601.

[3] M. M. Aziz, A. A. Bakar, and M. R. Yaakub, "CoreNLP dependency parsing and pattern identification for enhanced opinion mining in aspect-based sentiment analysis," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102035, Apr. 2024, doi: 10.1016/j.jksuci.2024.102035.

[4] S. O. Khairunnisa, Z. Chen, and M. Komachi, "Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 6, pp. 1–21, Jun. 2023, doi: 10.1145/3592854.

[5] Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, "Improving neural machine translation with POS-tag features for low-resource language pairs," *Heliyon*, vol. 8, no. 8, p. e10375, Aug. 2022, doi: 10.1016/j.heliyon.2022.e10375.

[6] N. Fatima, S. M. Daudpota, Z. Kastrati, A. S. Imran, S. Hassan, and N. S. Elmitwally, "Improving news headline text generation quality through frequent POS-Tag patterns analysis," *Eng. Appl. Artif. Intell.*, vol. 125, p. 106718, Oct. 2023, doi: 10.1016/j.engappai.2023.106718.

[7] S. Chotirat and P. Meesad, "Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning," *Heliyon*, vol. 7, no. 10, p. e08216, Oct. 2021, doi: 10.1016/j.heliyon.2021.e08216.

[8] A. Maulana and A. Romadhony, "Domain Adaptation for Part-of-Speech Tagging of Indonesian Text Using Affix Information," *Procedia Comput. Sci.*, vol. 179, pp. 640–647, 2021, doi: 10.1016/j.procs.2021.01.050.

[9] M. Kurniawan, K. Kusrini, and M. R. Arief, "Part of Speech Tagging Pada Teks Bahasa Indonesia dengan BiLSTM + CNN + CRF dan ELMo," *J. Eksplora Inform.*, vol. 11, no. 1, pp. 29–37, Jan. 2022, doi: 10.30864/eksplora.v11i1.506.

[10] D. E. Cahyani and W. Mustikaningtyas, "Indonesian part of speech tagging using maximum entropy markov model on Indonesian manually tagged corpus," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 11, no. 1, p. 336, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp336-344.

[11] M. Mursyit, A. P. Wibawa, I. A. E. Zaeni, and H. A. Rosyid, "Pelabelan Kelas Kata Bahasa Jawa Menggunakan Hidden Markov Model," *Mob. Forensics*, vol. 2, no. 2, pp. 71–83, Aug. 2020, doi: 10.12928/mf.v2i2.2450.

[12] R. A. Pratama, A. A. Suryani, and W. Maharani, "Part of Speech Tagging for Javanese Ngoko Language with Hidden Markov Model," vol. 4, no. 1, 2020.

[13] N. P. Dewi and U. Ubaidi, "POS Tagging Bahasa Madura dengan Menggunakan Algoritma Brill Tagger," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 7, no. 6, pp. 1121–1128, Dec. 2020, doi: 10.25126/jtiik.2020722449.

[14] I. Firmansyah, P. P. Adikara, and S. Adinugroho, "Klasifikasi Kelas Kata (Part-Of-Speech Tagging) untuk Bahasa Madura Menggunakan Algoritme Viterbi," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 8, no. 5, pp. 1039–1048, Oct. 2021, doi: 10.25126/jtiik.2021854483.

[15] A. Sumoko, A. B. P. Negara, and H. S. Pratiwi, "Perbandingan Tipe Metode PoS Tagger Terhadap Nilai Akurasi Untuk Bahasa Melayu Pontianak," *J. Sist. Dan Teknol. Inf. Justin*, vol. 9, no. 3, p. 342, Aug. 2021, doi: 10.26418/justin.v9i3.44116.

[16] S. Ullah *et al.*, "A Deep Learning-Based Approach for Part of Speech (PoS) Tagging in the Pashto Language," *IEEE Access*, vol. 12, pp. 86355–86364, 2024, doi: 10.1109/ACCESS.2024.3412175.

[17] A. Zilziana, A. A. Suryani, and I. Asror, "Part Of Speech Tagging Menggunakan Bahasa Jawa Dengan Metode Condition Random Fields".

Vol. 7, No. 3, June 2026, Page. 2534-2543
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5361

[18] D. Hoesen and A. Purwarianti, "Investigating Bi-LSTM and CRF with POS Tag Embedding for Indonesian Named Entity Tagger," in *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia: IEEE, Nov. 2018, pp. 35–38. doi: 10.1109/IALP.2018.8629158.
[19] 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, p. 10, Jan. 2022, doi: 10.1186/s40537-022-00561-y.
[20] A. A. Kha *et al.*, "Comparison of Machine Learning and Deep Learning Models for Part-of-Speech Tagging".
[21] M. Alfian, U. L. Yuhana, and D. Siahaan, "Indonesian Part-of-Speech Tagger: A Comparative Study," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Lombok, Indonesia: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICAICTA59291.2023.10390353.
[22] M. Kamayani, "Perkembangan Part-of-Speech Tagger Bahasa Indonesia," *J. Linguist. Komputasional JLK*, vol. 2, no. 2, p. 34, Sep. 2019, doi: 10.26418/jlk.v2i2.20.
[23] A. Benlahbib, A. Boumhidi, A. Fahfouh, and H. Alami, "Comparative Analysis of Traditional and Modern NLP Techniques on the CoLA Dataset: From POS Tagging to Large Language Models," *IEEE Open J. Comput. Soc.*, vol. 6, pp. 248–260, 2025, doi: 10.1109/OJCS.2025.3526712.
[24] W. Wongso, H. Lucky, and D. Suhartono, "Pre-trained transformer-based language models for Sundanese," *J. Big Data*, vol. 9, no. 1, p. 39, Dec. 2022, doi: 10.1186/s40537-022-00590-7.
[25] Y. Sudaryat, *Struktur bahasa Sunda: sintaksis dalam gamitan pragmatik*, Cetakan pertama. Bandung, Indonesia: UPI Press, 2019.
[26] D. Soyusiawaty and A. Fadlil, "Pengembangan Korpus Bahasa Minang pada Spell Error Corpus for Minang Language (SPEML)," vol. 11, no. 01, 2025.
[27] A. Sulastri1, "Geolinguistik: Variasi Dialek Dan Lemahnya Pemertahanan Bahasa Sunda Oleh Generasi Muda," *J. Geogr.*, vol. 13, no. 1, pp. 38–46, Oct. 2024, doi: 10.24036/geografi/vol13-iss1/3970.
[28] R. A. Danadibrata, *Kamus basa Sunda*, Cet. 1. Bandung: Wedalan Panitia Penerbitan Kamus Basa Sunda, gawe bareng PT Kiblat Buku Utama,jeung Universitas Padjadjaran, 2006.
[29] D. Koswara, "Racikan Sastra," *Bdg. Jur. Pendidik. Bhs. Drh. UPI*, 2013.
[30] A. Rosidi, *Babasan & paribasa: kabeungharan basa Sunda*. Kiblat Buku Utama, 2022.
[31] M.-C. De Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal Dependencies," *Comput. Linguist.*, pp. 1–54, May 2021, doi: 10.1162/coli_a_00402.
[32] Y. Sudaryat, A. Prawirasumantri, and K. Yudibrata, *Tata basa Sunda kiwari*, Cet. 1. Bandung: Yrama Widya, 2007.
[33] A. ARDIYANTI SURYANI, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "PoSTagged Sundanese Monolingual Corpus." Telkom University Dataverse, 2022. doi: 10.34820/FK2/VTAHRH.