Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 2, June, 2024

Comparative Analysis of Deep Learning Models for Part of Speech Tagging in the Malay Language

Bakare Mustaphaa Adebayo [1], Kalaiarasi Sonai Muthu Anbananthen [1]*, Saravanan Muthaiyah [2], Saravanan Nathan Lurudusamy [3]

[1] *Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[2] *School of Business and Technology, International Medical University, Kuala Lumpur 57000, Malaysia.*

[3] *Division Consulting & Technology Services, Telekom Malaysia, Kuala Lumpur 50672, Malaysia.*

Received 10 January 2024; Revised 18 April 2024; Accepted 03 May 2024; Published 01 June 2024

Abstract

Despite the widespread use of Malay, under-resourced languages like Malay face challenges in Natural Language Processing (NLP), particularly in Part-of-Speech (POS) tagging. The scarcity of annotated corpora poses a primary obstacle to POS tagging in Malay. This study aims to enhance the effectiveness and reliability of POS tagging models explicitly tailored for under-resourced languages within the field of NLP, focusing on Malay. Existing models, which rely on Conditional Random Fields and Hidden Markov Models, exhibit limitations, underscoring the need for more robust approaches. The research conducts a comparative analysis of various deep-learning models with different encoders for POS tagging in Malay sentences. The experimental analysis demonstrates that the Bidirectional Long Short-Term Memory (Bi-LSTM) model, leveraging a pre-trained Bidirectional Encoder Representations from Transformers (BERT) embedding model, achieves exceptional accuracy, precision, recall, and F1 scores in predicting tags. Notably, the BERT + Bi-LSTM model, boasting an accuracy of 98.82%, outperforms other models, showcasing superior performance across all evaluated metrics. Additionally, this combined model effectively handles known and unknown words, yielding highly accurate POS tagging results for Malay sentences.

Keywords: Part of Speech Tagging; Deep Learning; Malay Text; Malay POS Tagger.

1. Introduction

Part-of-speech (POS) tagging is a crucial research field under the umbrella of natural language processing (NLP). POS tagging involves assigning each word in a sentence with its corresponding part of speech tag, such as a noun, verb, or adjective [1, 2]. Developing an accurate model for POS tagging requires substantial linguistic expertise and a vast amount of annotated corpora. The significance of POS tagging extends across various NLP applications, including name entity recognition, machine translation, and sentiment analysis. NLP applications can be executed on several levels, such as words, phrases, sentences, or documents. Although computers cannot comprehend human languages the same way humans can, they can assist humans in processing massive amounts of linguistic data. As the data associated with natural language undergoes continuous expansion, humans' manual analysis and extracting relevant information [3] become increasingly challenging. Therefore, the need for computer assistance has become increasingly important. Consequently, natural language processing has emerged as an intriguing subject of study within the realm of information technology and allied fields.

* Corresponding author: kalaiarasi@mmu.edu.my

http://dx.doi.org/10.28991/HIJ-2024-05-02-04

Despite recent advancements in NLP, POS tagging remains challenging, particularly for under-resourced languages like Malay. The lack of annotated corpora is one of the main reasons why POS tagging for Malay is difficult [4]. Annotated corpora are crucial for supervised learning-based approaches, where the model is trained on labeled data to make predictions on new, unseen data. Unfortunately, no standard Malay corpus has been developed for POS tagging, making it difficult for researchers to obtain sufficient labeled data to train and evaluate their models [5]. While several POS tagging models have been proposed for Malay, they mainly employ sequential models, such as Hidden Markov Models (HMM) [6], Conditional Random Fields (CRF), along with traditional machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT), and K-Nearest Neighbors (KNN) [7]. These models have shown reasonable performance in POS tagging for Malay, but further enhancements can still be made.

To enhance the efficiency of POS tagging models in NLP applications, researchers have increasingly explored the application of deep learning approaches. In many NLP applications, deep learning techniques are more effective than other traditional model training methods. Deep learning approaches have gained widespread recognition and popularity in NLP due to advancements in processing power, hardware, and so on [8]. It has shown promising results in POS tagging for rich-sourced languages like English. However, deep learning applications in under-resourced languages like Malay remain limited. Some researchers have explored deep learning for POS tagging in Malay and other low-resource languages, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). For example, Tiun et al. [5] developed a Bidirectional LSTM-CRF deep learning model for Malay POS tagging, achieving a 0.94 F1 score.

Despite the progress in developing POS tagging models for Malay, much work needs to be done to develop a more reliable and robust model that can aid in advancing the field of NLP in this language. Therefore, the primary objective of this study is to explore how deep learning approaches can be utilized for POS tagging in Malay. By developing more accurate and robust models, we aim to contribute to advancing NLP applications, especially for under-resourced languages.

2. Literature Review

POS tagging means assigning labels that imply the word's linguistic class in the sentence. This process began with text annotation [9], in which human annotators manually assigned tags to each word in a text. Assigning tags to each text manually is very difficult and time-consuming. Therefore, researchers have developed several automated methods to perform this task. The rule-based method, the earliest and most widely utilized approach, includes manually building a set of rules based on training data. The application of the rule-based method in English was initially reported in prior studies [10, 11]. The efficacy of the rule-based approach in the English language resulted in its implementation in other languages, such as Hindi [12] and Korean [13]. The rule-based approach employs a predetermined set of linguistic rules for text tagging.

Furthermore, certain rule-based methodologies incorporate the utilization of a lexicon to enhance the precision of part-of-speech (POS) tagging. Previous researchers have employed regular expressions to formulate a collection of rules. Regular expressions detect and assign tags to specific patterns within the text. The rule-based strategy generally shows simple, understandable, and clear advantages [6]. The established rules are understandable and unambiguous. However, the rule-based approach has several issues. A significant challenge lies in generating the rules through a manual process. The entire process is time-consuming and susceptible to errors. Furthermore, it should be noted that the rules formulated for a specific data set may not be applicable or useful to another data set that falls outside the domain. Another challenge associated with rule-based methods is their difficulty effectively handling unknown words in the model.

Researchers have developed statistical methods, including HMMs, Maximum Entropy Markov Models (MEMMs), and CRFs, to overcome the constraints associated with the rule-based method. The statistical models utilize machine learning techniques to learn patterns within annotated corpora, subsequently using these learned patterns to analyze and interpret new textual data. The HMM models have been used for several English Language corpora, including the BROWN corpus, with accuracy from 76% to 95%. However, the accuracy of the models is usually proportional to the number of tokens used in the model training [14]. HMM models have also been used for other less common languages like Arabic [15], Azerbaijani [16], Indonesian [17], and Nepali [18], with accuracy spanning from 70% to 90%. A notable drawback of HMM models arises when encountering unknown words—those not encountered during training. In such cases, the model's accuracy tends to decrease significantly. CRF has been introduced to address the problem of unknown words. CRF captures more sequential dependencies and improves performance when predicting unknown words.

CRF models build upon the HMM method and offer a greater capacity to learn dependencies. Consequently, they have been utilized as a POS tagging model in low- and high-resource languages. In a comparative study of HMM and CRF models applied to the low-resource language Yoruba, Ayogu found that the CRF model had a slight advantage over the HMM model in terms of accuracy [19]. Other low-resource languages, such as Malayalam [20], Urdu [21], Assamese [22], and Vietnamese [23], have also utilized CRF models. High-resource languages, such as English, have also seen the application of CRF models [24, 25]. Although other machine learning approaches, such as the decision tree approach, have been used for POS tagging, HMM and CRF are the most commonly used approaches [26].

The MEMM model uses machine learning techniques to learn patterns in an annotated corpus and apply the learned patterns to new texts. Based on the maximum entropy principle, it ensures the probability distribution for a set of events is as uniform as possible, subject to constraints imposed by available information. The model has been effectively utilized in POS tagging for both English and Spanish [27, 28]. Ratnaparkhi [27] developed an English POS tagger using MEMM. This tagger demonstrated good performance on the Penn Treebank corpus. Taulé & Martí [28] developed a POS tagger for the Spanish language using a MEMM approach. Their system demonstrated high accuracy on large-scale Spanish text. MEMM models have been employed in various low-resource languages like Chinese, Arabic, and Korean. One notable benefit of MEMMs is their ability to handle a large range of dependencies in textual data efficiently. This feature is very useful in languages with complex grammatical structures. However, MEMMs faced criticism due to their long training times and large amounts of annotated data required to attain significant accuracy.

Recently, there has been an increasing inclination towards utilizing deep learning methodologies in POS tagging. Various approaches, including Recurrent Neural Networks (RNNs), CNNs, and Transformer models, have exhibited positive results in this domain. RNNs, particularly LSTM networks, have become increasingly popular due to their effectiveness in processing sequential data [29]. In terms of transformer models, the transfer learning ability improved the accuracy of many NLP applications [30, 31]. The study by Gopalakrishnan et al. [32] investigated the performance of LSTM and Gated Recurrent Unit (GRU) models on a biomedical dataset. The researchers found that the bi-directional versions of both LSTM and GRU models outperformed their respective simple models, showing superior results. The bi-directional LSTM model achieved the highest accuracy rate of 94.80%. However, using deep learning methodologies for POS tagging in languages with limited resources poses certain difficulties, primarily due to the limited availability of annotated datasets and the lack of experts during the development of these datasets. Despite these limitations, some studies have applied deep learning to POS tagging in languages such as Malayalam [33], Nepali [34], Bengali [35], Khasi [36], and Korean [37]. Further research in this area is ongoing.

Specifically, for the Malay language, recent studies using deep learning for POS tagging show promising results. Malay is spoken by over 300 million individuals worldwide, primarily in Singapore, Brunei, Indonesia, and Malaysia. This language belongs to the Austronesian language family and features a complex grammar structure that poses challenges for POS tagging. In recent years, there has been progress in using deep learning techniques to improve the accuracy of POS tagging in Malay. Tiun et al. [5] developed a method using a BiLSTM-CRF model to analyze Malay tweets. The model used Malay Tweet embeddings to convert the text into vector representations. These vectors were then fed into the BiLSTM, and the resulting output was used for classification with a CRF layer. Tiun et al. [5] compared this approach with SVM, NB, DT, and KNN. The evaluation of the BiLSTM-CRF model showed a 2% improvement in F1-score compared to SVM, reaching an impressive 94%. However, it's important to note that the model's ability to work well with new data may be limited due to the relatively small dataset used for training and testing (1,253 instances for training and 538 for testing). Additionally, tweets have short sentences because of the 280-character limit, resulting in limited words in each dataset.

Table 1 provides a brief overview of various POS tagging approaches. The results indicate that deep learning models have exhibited higher efficacy in POS tagging than the rule-based approach and statistical machine learning methods. Although the rule-based approach offers simplicity and interpretability, statistical-based methods such as HMMs and CRFs provide better accuracy and generalization capabilities. However, in the context of the Malay language, deep learning models are showing more promise for POS tagging. Nevertheless, further research is necessary to compare different deep learning-based models for POS tagging in the Malay language.

Table 1. Summary of existing POS tagging approach

Author	Method	Language	Dataset Source	Result
			Rule Based	
Brill (1992) [11]		English	Brown Corpus	Error rate = 7.9%
			Machine Learning	
Fanoon (2019) [24]	CRF	English	Gimpel et al. (2011) [38]	Accuracy = 72.00%
Zhang et al. (2009) [25]	CRF	English	PFR segment & POS tagging corpora on the people's daily in January 1998	Precision = 95.79%
Zhang et al. (2009) [25]	HMM	English	PFR segment & POS tagging corpora on the people's daily in January 1998	Precision = 92.53%
Marquez (1999) [26]	Decision Tree	Spanish	The Wall Street Journal Annotated Corpus	Overall Accuracy = 96.84% Known Accuracy = 97.21% Ambiguous Accuracy = 91.95% Unknown Accuracy = 80.70%
Tran et al. (2009) [23]	CRF	Vietnamese	Manually annotated corpus	Precision = 91.64%
Albared et al. (2010) [15]	HMM	Arabic	Manually annotated corpus	Accuracy = 95.80%
ArchanaTC et al. (2014) [20]	CRF	Malayalam	Manually annotated corpus	Accuracy = 86.70%
Paul et al. (2016) [18]	HMM	Nepali	Manually annotated corpus	Accuracy = 96%
Ayogu et al. (2017) [19]	CRF	Yoruba	Manually annotated corpus	Accuracy = 84.66%
Mammadov et al. (2018) [16]	HMM	Azerbaijani	Manually annotated corpus	Accuracy = 90.00%
Cahyani et al. (2019) [17]	HMM	Indonesian	Manually annotated corpus	Accuracy = 77.56%
Ranjan Deka et al. (2020) [22]	CRF	Assamese	Manually annotated corpus	Accuracy = 91%

Nasim et al. (2020) [21]	CRF	Urdu	Jawaid et al. (2014) [39]	Accuracy = 95.80%
				F1-Score = 96.00%

		Deep Learning		
Kabir et al. (2016) [36]	Neural Network	Bengali	LDC2010T16 and ISBN 1- 58563-561-8 corpus	Accuracy = 93.33%

Gopalakrishnan et al. (2019) [32]	Bi-LSTM	English	GENIA version 3.02	Accuracy = 94.80%
				Precision = 95.00%
				Recall = 95.00%
				F1-Score = 95.00%

Kumar et al. (2017) [33]	Bi-LSTM	Malayalam	Manually annotated corpus	Accuracy = 87.57%
				Precision = 87.48%
				Recall = 87.57%
				F1-Score = 87.39%

Sarbin et al. (2020) [34]	Bi-LSTM	Nepali	Madan Puraskar Pustakalaya	Accuracy = 97.27% Loss Value = 0.0190

Nasim et al. (2020) [21]	Bi-LSTM-CRF	Urdu	Jawaid et al. (2014) [39]	Accuracy = 96.30%
				F1-Score = 96.00%

Hoojon et al. (2023) [35]	Bi-LSTM-CRF	Khasi	Manually annotated corpus	Accuracy = 98.90%
				Precision = 99.00%
				Recall = 99.00%
				F1-Score = 99.00%

Tiun et al. (2022) [5]	Bi-LSTM-CRF	Malay	Manually annotated corpus	Precision = 94.00%
				F1-Score = 94.00%
				Recall = 94.00%

Song et al. (2020) [37]	Bi-LSTM-CRF	Korean	Manually annotated corpus	Accuracy = 95.28%
				F1-Score = 97.27%

3. Research Methodology

This section outlines the methodology and experimental procedure for developing a Malay language POS tagging model using deep learning. The first step involves creating a Malay Corpus to serve as the training data for the model. The training model architecture is then described, considering the hyperparameters required during training.

3.1. Malay Corpus Development

Creating a corpus to develop a POS tagging model was necessary since no standard corpus was available for the Malay language. This study manually created a corpus by collecting online newspaper articles from several sources, including Berita Harian, Harakah, and Kosmo, between August 2022 and February 2023. The manual annotation of the tags used the existing tagset definition of Mohamed et al. [6], consisting of 21 tags [6]. However, four tag definitions (KP, #E, @KG, SEN) are absent in the collected newspaper data sample. These tags were removed from the list of tags used in this training. In addition, two new tag definitions are added to the list of tags. First (NM) represents a word that cannot be found in the Malay Dictionary for tagging. Secondly (PP) represents a Malay phrase combining two or more words to form a single word. The total of 955 sentences in the corpus contains 98,832 words, 10,507 distinct words, and 1,252 ambiguous words.

3.2. Training Model Architecture

This study uses deep learning techniques to develop a POS tagging model for the Malay language by exploring different deep learning models and encoders. The models employed are based on prior POS tagging methods used in Malay and other languages, comprising LSTM, GRU, Bi-LSTM, and Bi-GRU. Furthermore, to effectively encode the dataset features, this study compares the utilization of a traditional One-Hot encoder with a BERT model, which has shown promising results in deep learning. The architecture of the training model is shown in Figure 1. Based on the design, 8 models are derived by combining different deep-learning layers and encoders. The architecture's first layer is the input containing the training corpus. The input is encoded using the One-Hot or the BERT encoder, and the vectorized output is passed to the second architecture layer, the deep learning layer. The deep learning layer comprises two layers. The first deep learning layer consists of a single layer of LSTM, GRU, Bi-LSTM, or Bi-GRU. The second output layer is a time-distributed dense layer with a SoftMax activation function (Figure 2) shows a simple example sequence of four tokens, demonstrating how the encoded input is generated using either the One-Hot encoder or the BERT and passed to the deep learning layer to generate the output tags.

LSTM is a type of RNN capable of capturing long-term dependencies in sequential data, such as text, speech, and time-series data. LSTMs achieve this by incorporating a memory cell and three gating mechanisms: input gate, forget gate, and output gate [40]. On the other hand, GRU is also a type of recurrent neural network like LSTM capable of capturing long-term dependencies in sequential data. However, it has a simpler architecture compared to LSTM. The

GRU has only two gates: a reset gate and an update gate [41]. Bi-LSTM and Bi-GRU are bidirectional variants of LSTM and GRU, where information flows in both directions. These architectural designs have successfully demonstrated their ability to capture context and dependencies in sequential data.

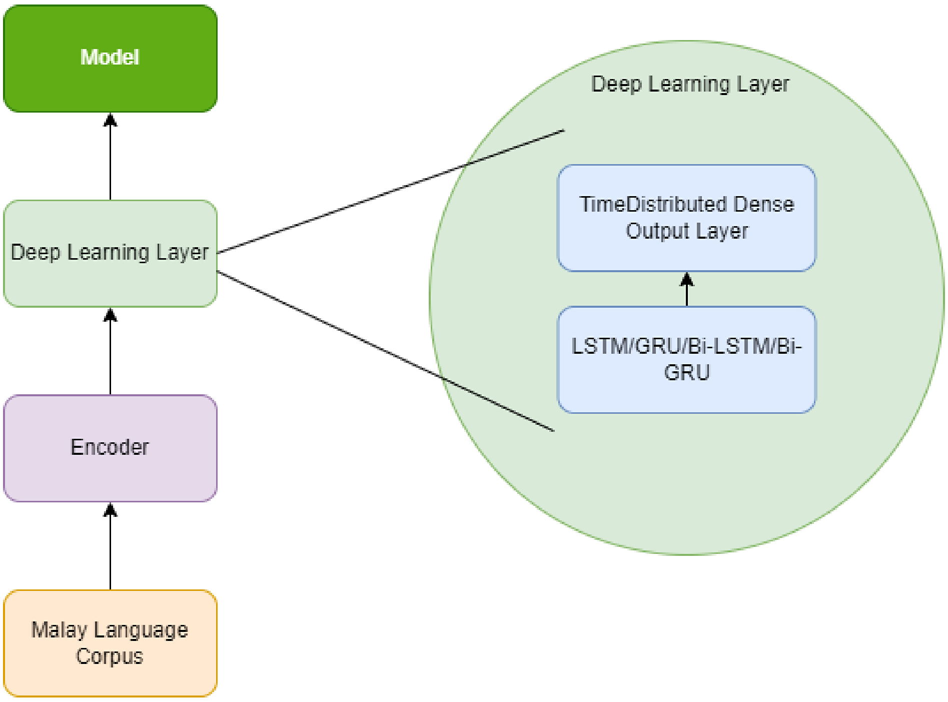

Figure 1. Training model architecture

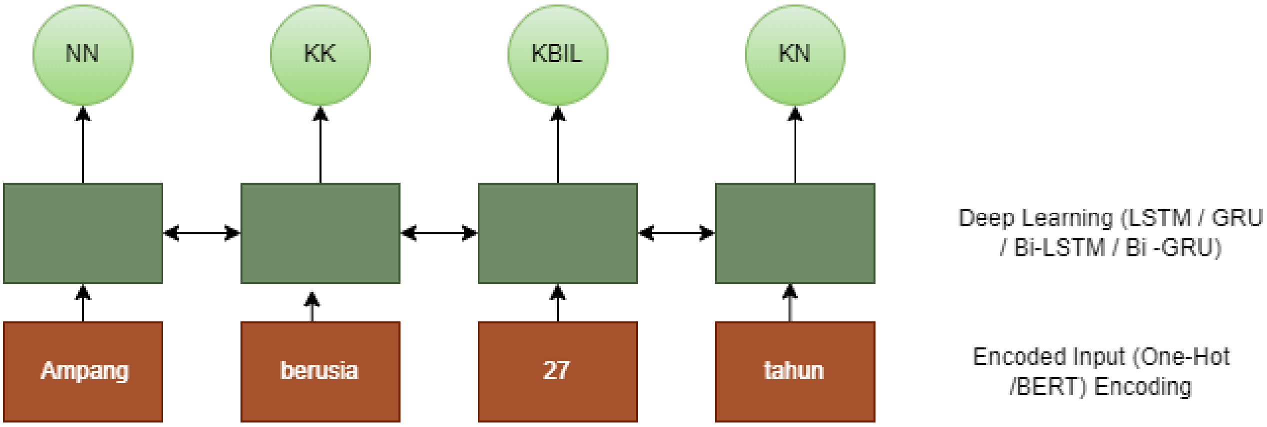
NN
KK
KBIL
KN
Ampang
berusia
27
tahun
Deep Learning (LSTM / GRU / Bi-LSTM / Bi -GRU)
Encoded Input (One-Hot /BERT) Encoding

Figure 2. A sample of the input sequence

3.3. Training Model Architecture

The dataset is divided into training and testing data to facilitate the training of the POS tagging model. This study employs deep learning models with data encoding, and the TensorFlow framework is utilized. To achieve optimal performance, hyperparameters are fine-tuned for each model. The values and optimal choices of various hyperparameters tested in the experiment are presented in Table 2. Based on accuracy, the optimized values for each model are determined and outlined in Table 3. Each model is trained using a maximum of 100 epochs. The sentences in the dataset are randomized, allocating 20% for testing and utilizing the remaining 80% for training.

Table 2. Possible hyperparameter configuration

Hyperparameters	Tested Values
Deep Learning Layer Unit	64, 128, 256, 512
Optimizer	Adam, SGD, RMSprop
Learning rate	0.001, 0.0001, 0.00001, 0.000001

Table 3. Optimal hyperparameters

Model Name	Layer Unit	Learning rate	Optimize
BERT + LSTM	256	0.0001	Adam
One-Hot + LSTM	128	0.0001	RMSProp
BERT + GRU	512	0.0001	Adam
One-Hot + GRU	128	0.0001	Adam
BERT + Bi-LSTM	512	0.001	RMSProp
One-Hot + Bi-LSTM	128	0.001	RMSProp
BERT + Bi-GRU	128	0.0001	RMSProp
One-Hot + Bi-GRU	64	0.0001	Adam

4. Experimental Result

This section comprehensively analyses the proposed model for tagging Malay sentences based on their accuracy, precision, recall, and F1 scores.

Table 4 and Figure 3 show the accuracy, precision, recall, and F1-score for different encoders and models tagging Malay sentences. The experimental results indicate that the models using BERT encoding consistently outperform the models using One-Hot encoding in all evaluated metrics, as shown in Figure 3. This results from BERT's ability to capture conceptual meanings from words and leverage this information for training. In terms of model performance, The BERT + Bi-LSTM model outperforms the other models and achieves the highest performance across all metrics. It attains an accuracy of 98.82%, indicating its ability to correctly classify the POS tags for Malay sentences. Furthermore, its precision, recall, and F1-score of 0.98, 0.97, and 0.98 demonstrate that POS tags can be accurately identified and classified.

The model's performance in identifying known and unknown words is evaluated using the BERT encoder + Bi-LSTM model, shown in Table 5. We tested 9,260 words, of which 8,449 were known to the model in its training phase and 811 were unknown. Overall, the model achieves an accuracy of 98.82% in identifying POS tags for a wide range of Malay words. The model accurately tags known words encountered during training with an accuracy rate of 98.92%. This indicates that the model can effectively detect familiar word patterns and linguistic characteristics. In addition, the model predicts unknown words that are not included in the training data, with a commendable accuracy of 96.55%. This implies that the model can robustly generalize its knowledge of the Malay language and assign POS tags to unknown words based on that knowledge.

Table 4. Experimental results of models

Model	Accuracy (%)	Precision	Recall	F1-Score
BERT + LSTM	96.00	0.96	0.94	0.95
One-Hot + LSTM	94.24	0.94	0.93	0.94
BERT + GRU	96.04	0.95	0.95	0.95
One-Hot + GRU	94.16	0.93	0.95	0.94
BERT + Bi-LSTM	**98.82**	0.98	0.97	0.98
One-Hot + Bi-LSTM	94.61	0.94	0.94	0.94
BERT + Bi-GRU	97.63	0.98	0.96	0.97
One-Hot + Bi-GRU	95.51	0.95	0.93	0.94

Table 5. Experimental evaluation of known and unknown words

Model	Overall Testing Accuracy	Accuracy of known words	Accuracy of Unknown Words
BERT+Bi-LSTM	98.82%	98.92%	96.55%

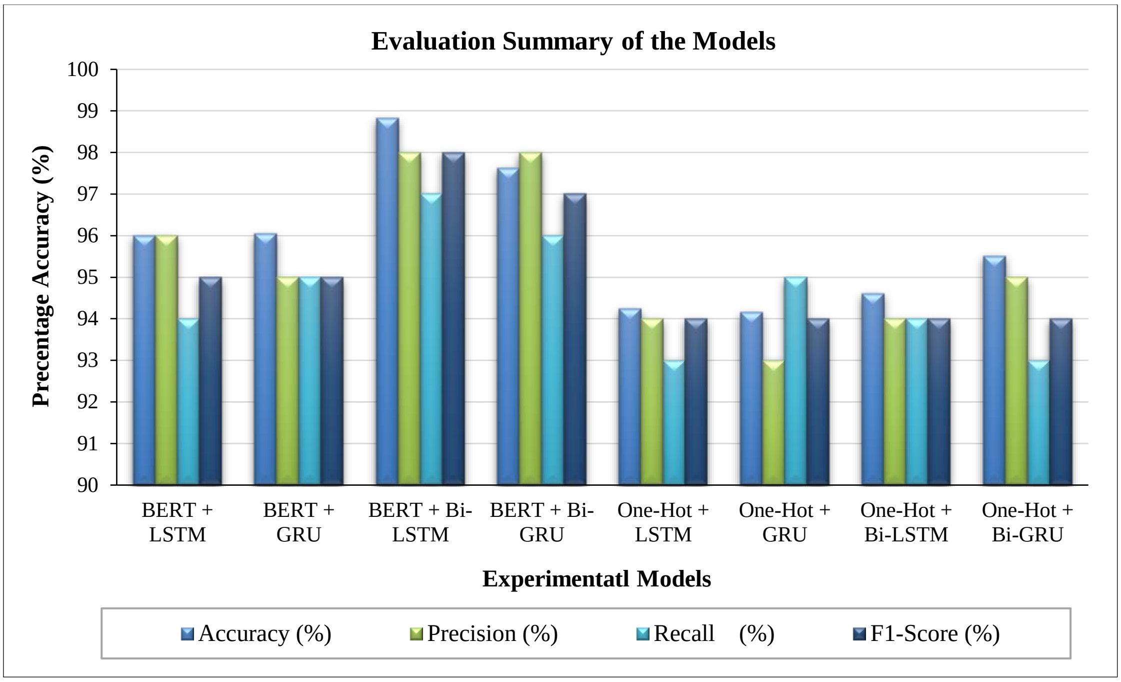

Figure 3. Evaluation summary of models

4.1. Discussion

This study examined various deep-learning models with different encoders and assessed their performance on known and unknown words. Among them, the BERT + Bi-LSTM model stood out for its remarkable ability to grasp the complex patterns and connections within the Malay language. Bi-LSTM effectively captures the intricate language patterns, resulting in precise and accurate POS labelling.

In contrast, the One-Hot plus GRU model performed the worst, with a 94.16% accuracy rate. Despite displaying a reasonable level of accuracy, this model trailed behind the competition. The lower accuracy, precision, recall, and F1-score of the One-Hot + GRU model indicate that it has difficulty capturing the intricate Malay language patterns and dependencies. This limitation may be attributed to one-hot encoding, which fails to capture the requisite semantic and contextual information for accurate POS tagging. In addition, One-Hot encoding represents each word as an independent binary vector, resulting in a very high-dimensional sparse vector. On the other hand, BERT has lower-dimensional dense representations, capturing more information in a compact form. Also, BERT + Bi-LSTM effectively handles known and unknown words, resulting in highly accurate POS tagging outcomes for Malay sentences.

Tiun et al. [5] introduced the only deep learning-based model for POS tagging in Malay based on the literature review. Our best model, combining BERT and Bi-LSTM, was compared to Tiun et al.'s model, which used the Bi-LSTM + CRF algorithm. Table 6 shows that Tiun's model achieved precision, recall, and F1 scores of 94%, while our proposed model achieved higher scores of 98%, 97%, and 98% for precision, recall, and F1, respectively. We used a similar dataset to the one employed by Tiun et al., indirectly allowing for a comparison between the models based on data.

Table 6. Model comparison

Model	Precision	Recall	F1-Score
Bi-LSTM+CRF [5]	0.94	0.94	0.94
BERT + Bi-LSTM (Proposed Model)	0.98	0.97	0.98

The superior performance of the proposed BERT encoder + Bi-LSTM model compared to the Bi-LSTM + CRF algorithm in Malay POS tagging can be attributed firstly, BERT provides contextualized word representations, enabling a better understanding of word dependencies within a sentence and capturing more nuanced contextual information. This contextual awareness is crucial for accurate POS tagging. Secondly, the combination of BERT with Bi-LSTM allows for robust feature extraction. BERT encodes the contextual word representations, while the subsequent Bi-LSTM layer refines these representations by capturing sequential patterns in the word order, resulting in a more comprehensive and informative sentence representation.

Furthermore, when used with Bi-LSTM, the CRF algorithm explicitly models the relationships between different POS tags. However, it can exhibit a bias towards more common tags while neglecting rarer ones, potentially leading to imbalanced predictions [42]. By employing the BERT encoder, the model is less susceptible to such biases, can provide more balanced predictions, and can better handle the full spectrum of POS tags.

5. Conclusion

This study explores the application of deep-learning models with different encoders for POS tagging in Malay sentences, specifically comparing BERT and One-Hot encoding models. The results consistently demonstrate that BERT encoding models outperform One-Hot encoding models, highlighting the effectiveness of contextualized word embeddings in capturing the nuances of the Malay language. Notably, the BERT + Bi-LSTM model achieves the highest accuracy of 98.82% among the evaluated models, showcasing its exceptional performance in comprehending complex patterns and dependencies in the language through the combined strengths of BERT and Bi-LSTM. These findings represent significant progress in natural language processing methodologies, offering valuable insights for advancing Malay language analysis. Future research will delve deeper into nuanced text analysis aspects, particularly by leveraging POS tagging to extract and categorize various linguistic elements. For example, researchers could explore how different parts of speech, such as nouns, adjectives, and verbs, contribute to the overall sentiment conveyed in a text. By mapping these aspects to sentiment analysis, we can better understand the sentiments and opinions expressed within Malay texts. Moreover, future research could explore the intersection of aspect-based sentiment analysis and entity recognition in Malay language texts. In a product review, entities such as brand names or product features may play a crucial role in shaping the sentiment expressed by the reviewer.

6. Declarations

6.1. Author Contributions

Conceptualization, A.M.B. and K.S.M.A.; methodology, A.M.B.; validation, S.M. and S.N.L.; formal analysis, A.M.B.; investigation, S.N.L.; writing—original draft preparation, A.M.B. and K.S.M.A.; writing—review and editing, S.M. and S.N.L.; supervision, K.S.M.A. and S.M.; funding acquisition, K.S.M.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available in the article.

6.3. Funding

This work was supported by the Multimedia University, Malaysia.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. Journal of Big Data, 9(1). doi:10.1186/s40537-022-00561-y.

[2] Anbananthen, K. S. M., Krishnan, J. K., Sayeed, M. S., & Muniapan, P. (2017). Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text. American Journal of Applied Sciences, 14(9), 843–851. doi:10.3844/ajassp.2017.843.851.

[3] Anbananthen, S. K., Sainarayanan, G., Chekima, A., & Teo, J. (2006). Data mining using pruned artificial neural network tree (ANNT). 2nd International Conference on Information & Communication Technologies, Damascus, Syria. doi:10.1109/ICTTA.2006.1684577.

[4] Ali, N. M., Ngo, G. H., & Lan, A. L. H. (2023). Construction of Part of Speech Tagger for Malay Language: A Review. Proceedings - 2023 5th International Conference on Natural Language Processing, (ICNLP 2023), 253–257. doi:10.1109/ICNLP58431.2023.00053.

[5] Tiun, S., Ariffin, S. N. A. N., & Chew, Y. D. (2022). POS Tagging Model for Malay Tweets Using New POS Tagset and BiLTSM-CRF Approach. CEUR Workshop Proceedings, 3315, 160–165.

[6] Mohamed, H., Omar, N., & Ab Aziz, M. J. (2011). Statistical Malay part-of-speech (POS) tagger using Hidden Markov approach. 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011, June, 231–236. doi:10.1109/STAIR.2011.5995794.

[7] Ariffin, S. N. A. N., & Tiun, S. (2018). Part-of-speech tagger for Malay social media texts. GEMA Online Journal of Language Studies, 18(4), 124–142. doi:10.17576/gema-2018-1804-09.

[8] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 53. doi:10.1186/s40537-021-00444-8.

[9] Sonai, K., Anbananthen, M., Mohamed, A., & Elyasir, H. (2013). Evolution of Opinion Mining. Australian Journal of Basic and Applied Sciences, 7(6), 359–370.

[10] Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. Computational Linguistics, 21(4), 543–565.

[11] Brill, E. (1992). A simple rule-based part of speech tagger. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 152-159. doi:10.3115/974499.974526.

[12] Garg, N., Goyal, V., & Preet, S. (2012). Rule-Based Hindi Part of Speech Tagger. International Conference on Computational Linguistics, 2(December), 163–174.

[13] Lee, G. G., Cha, J., & Lee, J. H. (2002). Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean. Computational Linguistics, 28(1), 53–70. doi:10.1162/089120102317341774.

[14] Tyagi, S., & Shankar Mishra, G. (2016). Statistical Analysis of Part of Speech (Pos) Tagging Algorithms for English Corpus. International Journal of Advance Research, Ideas and Innovations in Technology, 2(3), 1-9.

[15] Albared, M., Omar, N., Aziz, M. J. A., & Ahmad Nazri, M. Z. (2010). Automatic part of speech tagging for Arabic: An experiment using bigram hidden markov model. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6401 LNAI, 361–370. doi:10.1007/978-3-642-16248-0_52.

[16] Mammadov, S., Rustamov, S., Mustafali, A., Sadigov, Z., Mollayev, R., & Mammadov, Z. (2018). Part-of-Speech Tagging for Azerbaijani Language. IEEE 12th International Conference on Application of Information and Communication Technologies, AICT 2018 – Proceedings, Almaty, Kazakhstan. doi:10.1109/ICAICT.2018.8747154.

[17] Cahyani, D. E., & Vindiyanto, M. J. (2019). Indonesian part of speech tagging using hidden markov model - Ngram viterbi. 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019, 353–358. doi:10.1109/ICITISEE48480.2019.9003989.

[18] Paul, A., Purkayastha, B. S., & Sarkar, S. (2016). Hidden Markov Model based Part of Speech Tagging for Nepali language. 2015 International Symposium on Advanced Computing and Communication, ISACC 2015, 149–156. doi:10.1109/ISACC.2015.7377332.

[19] Ayogu, I. I., Adetunmbi, A. O., Ojokoh, B. A., & Oluwadare, S. A. (2017). A comparative study of hidden Markov model and conditional random fields on a Yorùba part-of-speech tagging task. Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, 2017-January, 1–6. doi:10.1109/ICCNI.2017.8123784.

[20] Krishnapriya, V., Sreesha, P., Harithalakshmi, T. R., Archana, T. C., & Vettath, J. N. (2003). Design of a POS tagger using conditional random fields for Malayalam. 2014 1st International Conference on Computational Systems and Communications, ICCSC 2014, 370–373. doi:10.1109/COMPSC.2014.7032680.

[21] Nasim, Z., Abidi, S., & Haider, S. (2020). Modeling POS Tagging for the Urdu Language. 2020 International Conference on Emerging Trends in Smart Technologies, ICETST 2020, Karachi, Pakistan. doi:10.1109/ICETST49965.2020.9080721.

[22] Deka, R. R., Kalita, S., Kashyap, K., Bhuyan, M. P., & Sarma, S. K. (2020). A Study of T'nT and CRF Based Approach for POS tagging in assamese language. In Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, 600–604. doi:10.1109/ICISS49785.2020.9315939.

[23] Tran, O. T., Le, C. A., Ha, T. Q., & Le, Q. H. (2009). An experimental study on Vietnamese POS tagging. 2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009, 23–27. doi:10.1109/IALP.2009.14.

[24] Fanoon, A. R. F. S., & Uwanthika, G. A. I. (2019). Part of speech tagging for Twitter conversations using Conditional Random Fields model. Proceedings - IEEE International Research Conference on Smart Computing and Systems Engineering, SCSE 2019, 108–112. doi:10.23919/SCSE.2019.8842669.

[25] Zhang, X., Huang, H., & Zhang, L. (2009). The application of CRFs in part-of-speech tagging. International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2009, Vol. 2, 347–350. doi:10.1109/IHMSC.2009.210.

[26] Marquez, Ll. (1999). Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees. Universitat Politècnica de Catalunya, Barcelona, Spain. doi:10.5821/dissertation-2117-93974.

[27] Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 1996, 133–142.

[28] Taulé, M., M.A. Martí, M. R. (2008). Ancora: Multilingual and multilevel annotated corpora. Proceedings of 6th International Conference on Language Resources and Evaluation, 96–101.

[29] Busst, M. M. A., Anbananthen, K. S. M., Kannan, S., Krishnan, J., & Subbiah, S. (2024). Ensemble BiLSTM: A Novel Approach for Aspect Extraction From Online Text. IEEE Access, 12(January), 3528–3539. doi:10.1109/ACCESS.2023.3349203.

[30] Bakare, A. M., Anbananthen, K. S. M., Muthaiyah, S., Krishnan, J., & Kannan, S. (2023). Punctuation Restoration with Transformer Model on Social Media Data. Applied Sciences (Switzerland), 13(3), 1685. doi:10.3390/app13031685.

[31] Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. HighTech and Innovation Journal, 4(2), 453–462. doi:10.28991/HIJ-2023-04-02-015.

[32] Gopalakrishnan, A., Soman, K. P., & Premjith, B. (2019). Part-of-Speech Tagger for Biomedical Domain Using Deep Neural Network Architecture. 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, 6–10. doi:10.1109/ICCCNT45670.2019.8944559.

[33] Kumar, S., Kumar, M. A., & Soman, K. P. (2019). Deep learning-based part-of-speech tagging for Malayalam twitter data (Special Issue: Deep learning techniques for natural language processing). Journal of Intelligent Systems, 28(3), 423–435. doi:10.1515/jisys-2017-0520.

[34] Sayami, S., & Shakya, S. (2020). Nepali POS Tagging Using Deep Learning Approaches. International Journal of Science, 17(2), 69–84.

[35] Hoojon, R., & Nath, D. A. (2023). BiLSTM with CRF Part-of-Speech Tagging for Khasi language. 2023 4th International Conference on Computing and Communication Systems, I3CS 2023, I3CS 2023, 1–7. doi:10.1109/I3CS58314.2023.10127278.

[36] Kabir, M. F., Abdullah-Al-Mamun, K., & Huda, M. N. (2016). Deep learning-based parts of speech tagger for Bengali. 2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016, 26–29. doi:10.1109/ICIEV.2016.7760098.

[37] Song, H. J., & Park, S. B. (2020). Korean part-of-speech tagging based on morpheme generation. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(3), 1–10. doi:10.1145/3373608.

[38] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2, 42–47.

[39] Jawaid, B., Kamran, A., & Bojar, O. (2014). A tagged corpus and a tagger for Urdu. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2938–2943.

[40] Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 2545–2568. doi:10.18653/v1/2021.naacl-main.201.

[41] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 1-9. doi:10.48550/arXiv.1412.3555.

[42] Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, 32.