Part-of-Speech (POS) Tagging for Standard Brunei Malay: A Probabilistic and Neural-Based Approach

Izzati Mohaimin [1,*], Rosyzie A. Apong [1], and Ashrol R. Damit [2]

[1] School of Digital Science, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam;
Email: rosyzie.apong@ubd.edu.bn (R.A.A.)
[2] Faculty of Arts and Social Sciences, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam;
Email: rahimy.damit@ubd.edu.bn (A.R.D.)
*Correspondence: 20m2052@ubd.edu.bn (I.M.)

***Abstract*—As online information increases over the years, text mining researchers developed Natural Language Processing tools to extract relevant and useful information from textual data such as online news articles. The Malay language is widely spoken, especially in the Southeast Asian region, but there is a lack of Natural Language Processing (NLP) tools such as Malay corpora and Part-of-Speech (POS) taggers. Existing NLP tools are mainly based on Standard Malay of Malaysia and Indonesian language, but there is none for the Bruneian Malay. We addressed this issue by designing a Standard Brunei Malay corpus consisting of over 114,000 lexical tokens, annotated using 17 Malay POS tagsets. Furthermore, we implemented two commonly used POS tagging techniques, Conditional Random Field (CRF) and Bi-directional Long Short-Term Memory (BLSTM), to develop Bruneian POS taggers and compared their performances. The results showed that both CRF and BLSTM models performed well in predicting POS tags on Bruneian texts. However, CRF models outperform BLSTM, where CRF using all features achieved an F-Measure of 92.06% on news articles and 90.71% of F-Measure on crime articles. Adding a batch normalization layer to the BLSTM model architecture increased the performance by 7.13%. To further improve the BLSTM models, we suggested increasing the training data and experimenting with different hyperparameter settings. The findings also indicated that modelling BLSTM with fastText has improved the POS prediction of Bruneian words.**

***Keywords*—part-of-speech tagging, Conditional Random Field (CRF), Bi-directional Long Short-Term Memory (BLSTM), pre-trained word embeddings, batch normalization**

I. Introduction

Malay language, or Standard Malay, is the official language of Brunei, Malaysia, Singapore, and Indonesia and has approximately 200 million speakers worldwide [1, 2]. Malaysia refer their standard Malay as Bahasa Malaysia, whereas Bahasa Indonesia is more common in Indonesia. However, there exist differences in pronunciation and vocabulary between the Malay language used in Brunei, Malaysia, and Singapore to the Malay language used in Indonesia [3, 4]. Although Standard Malay used in Brunei and Malaysia is closely related, there are some differences in grammatical syntax and semantics [5]. Some words that exist in Malay and Indonesian languages can occur in the same contexts, but the meaning may differ [4]. Therefore, ensuring a fair comparison between the works is difficult, and using their tagging models may not be accurate on Bruneian Malay texts.

Part-of-Speech (POS) tagging is a process of annotating every word in a sentence with a tag such as a noun, adjective, or verb based on the context and word definition. It is an essential process in the Natural Language Processing (NLP) pipeline, where the POS tags are used as inputs to other higher-level NLP tasks such as Named-Entity Recognition (NER) and Machine Translation [6–8].

The Malay language is widely spoken, especially in the Southeast Asian region, but there is a lack of NLP tools such as Malay corpora and POS taggers. Existing NLP tools are mainly based on Standard Malay of Malaysia and Indonesian language, but there is none for the Bruneian Malay.

The contribution of this paper includes designing a Standard Brunei Malay corpus consisting of over 114,000 lexical tokens, annotated using 17 Malay POS tags. The paper also explored two commonly used POS tagging techniques, Conditional Random Field (CRF) and Bi-directional Long Short-Term Memory (BLSTM), to develop Bruneian POS taggers. The experimental results of the models are compared and discussed to determine the suitable classifier for Bruneian texts. Finally, the paper shares the directions for future research to improve the current work.

The rest of this paper is organized as follows. Section II reviews existing POS taggers in Malay languages. Next, Section III describes the data collection process and the design of the Malay POS tagset used for the corpus annotation. Then, Section IV shows the experiments and

Manuscript received October 8, 2022; revised December 28, 2022; accepted January 30, 2023; published August 17, 2023.

modelling combinations for implementing the Bruneian POS tagger models using CRF and Deep Learning, specifically BLSTM. The performance of all models and the interpretations of the results are reported in Section V. Lastly, we summarize our work and recommend plans for future works in Section VI.

II. Related Work

Rule-based [9] and TnT tagger [10] were among the first POS taggers developed for the English language. The models achieved at least 95% and 96% of tagging accuracy, respectively. Conditional Random Field or CRF [11] is another widely used method, a probabilistic technique for sequence labelling. Previous research has shown that CRF-based models have been trained and evaluated on various languages such as Arabic [12, 13], French [14], and Indonesian [8, 15–17]. Several features have been applied and reported to increase the tagging accuracy of these models compared to using only the basic features such as unigrams and bigrams. In the current years, neural network-based POS taggers have been developed and have proved to obtain high POS tagging accuracy in many languages [8, 18–21].

For the Malay language, various techniques have been used for sequential labelling. Mohamed *et al.* [22] developed a Malay POS tagger using a Trigram Hidden Markov Model (Trigram HMM). Their model achieved 67.9% of accuracy for unknown words by only using prefixes (the first three letters of a word). In their work, they experimented with using prefixes and suffixes to predict the POS tags of words. The model is trained and evaluated on a corpus of 18,400 tokens, where they divided the data into 90% for training and 10% for testing. A total of 21 tags were used in their corpus, most of which are taken from a bilingual Malay-English dictionary. They replaced and added a few of their own POS tags, such as "SEN" to tag number list and "SYM" to tag symbols and punctuations.

Alfred *et al*. [7] developed RPOS, a rule-based POS tagger evaluated on news and biomedical articles. The accuracies are 89% and 86%, respectively. In their work, they built a POS tag dictionary from Thesaurus Bahasa Melayu, which consists of about 8,700 tagged words. They proposed using affixes and word relation rules taken from the same thesaurus. A word in a sentence will be tagged if the word exists in the POS tag dictionary and only have a single tag. If multiple tags exist for the word, relation rules will be applied, and a suitable POS tag will be selected. However, if the word does not exist in the POS tag dictionary, affixing rules will be applied, creating a new word and meaning, then applied word relation rules to determine its POS tag. The limitation of their model was that they could not predict the correct POS tags of English-borrowed words. They concluded that the performance of RPOS can be improved by having more word relations and POS tags in the POS tag dictionary. Halid and Omar [23] also used a rule-based technique for their POS tagger in which they used the same POS tags as Alfred *et al*. [7] and introduced an additional of 15 new POS tags and two additional word relation rules. The average performance achieved was 93.06%. In contrast, when using only the tags and word relation rules stated by Alfred *et al.* [7], the model achieved an average of 77.17% of tagging accuracy. Hamzah and Syed [24] also applied rule-based for POS tagging. They collected Malay texts from police reports, including daily reports and common texts as their corpus. The corpus is then annotated using four basic tags (noun, verb, adverb, and adjective). The author emphasized the importance of rules arrangement for tagging so the best performance can be achieved by manually reviewing and experimenting with different ordering of the rules. Their POS tagger achieved 88.4% tagging accuracy based on using only morphological knowledge.

Xian *et al*. [25] developed Mi-POS, a machine learning Malay POS tagger based on a probabilistic method, Maximum Entropy. They compared its performance with Trigram HMM [22], Lazy Man's tagger, an unsupervised tagger [26], and RPOS [7]. Mi-POS outperforms the other taggers with a tagging accuracy of 95.16% for news articles and 81.12% for non-news articles. Unlike other taggers used for the comparison, Mi-POS does not require additional dictionaries or translators and only uses basic probabilistic calculations, reducing overall processing time and high tagging accuracy. However, their model may be prone to genre bias due to corpus limitation, which can be avoided by including a wide variety of topics for training, such as medical articles and social media texts.

Tan *et al*. [21] implemented Long Short-Term Memory (LSTM), HMM, and Weighted Finite-State Transducers (WFST) for Malay POS tagging and compared their performances. WFST outperforms HMM and LSTM in terms of tagging performance when they use morphological information. However, LSTM can perform equally well with WFST when the morphological aspect is excluded. WFST also took the shortest time to train and decode, followed by HMM, then LSTM. The authors stated that LSTM networks could benefit languages with few and limited linguistic resources.

Malay POS taggers can still be improved and experimented with using techniques such as CRF and Deep Learning (DL) which are rarely used or none for the Malay language, specifically the Malay used by Brunei, Singapore, and Malaysia. In this paper, the two methods are used to develop Malay POS taggers and evaluated on Standard Brunei Malay corpus. Table I shows common POS tags across two works [7, 25] that will be used to annotate the Bruneian corpus.

TABLE I. Ten Common Tags Across Two Tagsets

Alfred *et al.* [7]	Xian *et al*. [25]
CC (Conjunction)	CC (Coordinate Conjunction)
CD (Cardinal Number)	CD (Cardinals)
IN (Preposition)	IN (Preposition)
JJ (Adjective)	JJ (Adjectives)
NEG (Negation)	NEG (Negations)
NN (Noun)	NN (Nouns)
NNP (Proper Noun)	NNP (Proper Nouns)
RB (Adverb)	RB (Adverbs)
VB (Verb)	VB (Verbs)
WP (Interrogative)	WH (WH)

III. Corpus Development

A. Malay POS Tagset

The main principle in designing a Malay POS tagset is simplicity [15, 27]. The purpose is to avoid cognitive overload on the annotators since the corpus is required to be manually labelled and checked while being able to preserve basic tags that can distinguish grammatical properties of words. There is currently no research on Bruneian tagset, hence other studies that have used Malay POS tagsets are reviewed and studied. Due to the similarities of the Malay language between Brunei and Malaysia, tagsets used for Standard Malaysian texts are chosen (see Table I). Another seven additional tags were introduced to solve tag ambiguity, which will be the final version of the Malay POS tagset for Brunei corpus annotation (see Table II).

Unlike other works which use one tag for any borrowed words from foreign languages [15, 27], we used the tag "X(KP)" to distinguish locally used common foreign words such as "O-levels" and "ta'zir". Other foreign words (e.g., "police" and "revised") and website links (e.g., "www.baiduri.com") are tagged as "X".

TABLE II. Final Version of Malay POS Tagset

Tag	Used by other works	Description (English)	Description (Malay)	Examples
AUX	[7]	Auxiliary	Kata Bantu	adalah, akan, telah
CC	[7, 25]	Conjunction	Kata Hubung	dan, serta, selain
CD	[7, 25]	Cardinal number	Kata Bilangan	beberapa, 2020, lima
DET	[25]	Determiner	Kata Penentu	ini, itu
IN	[7, 25]	Preposition	Kata Sendi	sini, dalam, pada
JJ	[7, 25]	Adjective	Kata Sifat	baharu, kecil, lebih
NEG	[7, 25]	Negation	Kata Nafi	tidak, bukan, jangan
NN	[7, 25]	Noun	Kata Nama	wang, hasil, teknologi
NNP	[7, 25]	Proper Noun	Kata Nama Khas	Tutong, MPK, Jumaat
PRP	[23, 25]	Pronoun	Kata Ganti Nama	ia, Kami, kitani
PT	-	Punctuation	Tanda Baca	., -, ,
RB	[7, 25]	Adverb	Kata Adverba	Sebagai, kemudiannya, lagi
SYM	[23]	Symbol	Simbol	%, /, =
VB	[7, 25]	Verb	Kata Kerja	meningkat, deijemput, membangun
WP	[7, 25]	Interrogative (WH Questions)	Kata Tanya	mana, apakah, siapakah
X	-	Others	Lain-lain	revised, kan, www
X(KP)	-	Others (Borrowed Words)	Kata Pinjam	O-levels, A-Levels

B. Data Collection

A Bruneian corpus is developed by collecting and crawling news articles from two sources: 1) BruDirect [28], a local online news website, and 2) a police government website [29], as shown in Fig. 1. BruDirect generally covers eight categories of news: National, Borneo, Southeast Asia, World, Business, Entertainment, Science and Technology, Health and Lifestyle. The government website publishes local crime news and general news about the department. These websites were chosen because they provide the official latest news in digital form, cover various topics, and are freely available in the Standard Malay language. Altogether, the corpus consists of over 3,000 sentences or 114,000 lexical tokens, which are annotated using the Malay POS tagset.

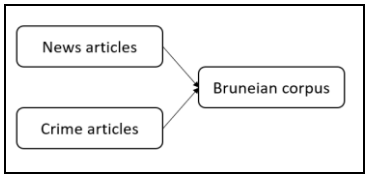

Figure 1. Bruneian corpus containing from two different sources (news from BruDirect website and crime articles from police website) to become one single corpus.

C. Data Cleaning

Data or text cleaning is a process to detect and fix any inaccurate data after scraping. This process is vital to ensure text analytics can use the data. The process involves general data cleaning, sentence splitting, and removing unwanted texts from the corpus. General cleaning includes removing any extra white spacing between tokens and extra new line spacing between sentences.

One important step for sentence splitting is called Sentence Boundary Detection (SBD) where it involves deciding the beginning and ending of a sentence. This process is critical because the learning model requires input sequences to learn the patterns and the sequence needs to be correct (in complete sentence). We use Punkt sentence tokenizer provided by Natural Language Toolkit (NLTK) because the default tokenizer cannot detect the correct boundaries between some Malay texts. The Punkt sentence tokenizer can also be trained on other languages not supported by the NLTK tokenizer, such as the Malay language. New abbreviations can also be added to the tokenizer to detect sentence boundaries accurately. We feed two inputs to the model to learn. The first input is a corpus containing the sentences where they are not split. The second input is a collection of abbreviations not usually present in English texts or not included in the abbreviation list in the NLTK sentence tokenizer. Table III lists the added abbreviations where most of the added abbreviations are different short forms of Muhammad, a common Malay name. It is important to note that the abbreviations need to be in lower case in the Punkt tokenizer, but in the actual corpus, the abbreviations are in upper case.

TABLE III. ADDED ABBREVIATIONS TRAINED ON PUNKT TOKENIZER

Abbreviation	Description	Example
Abd	Abdul	Dr. Hajah Mawarni binti Haji ***Abd.*** Hamid
Ar	Architect	Yang Berhormat Fdr. ***Ar.*** Dayang Siti Rozaimeriyanty binti Dato Seri Laila Jasa Haji Abdul Rahman
Bhd	Berhad	KSSUP telah melantik Muara Port Company Sdn. ***Bhd.*** (MPC), sebuah syarikat GLCs
Fdr	Doctor Fellow	Yang Berhormat ***Fdr.*** Ar. Dayang Siti Rozaimeriyanty mengajukan soalan
Ir	Engineer	Pengarah Urusan AlamSejagat Consulting Engineers, ***Ir.*** Affandy bin Mohd. Morshidi
Md	Mohammad	Surah Al-Fatihah yang dipimpin oleh Awang ***Md.*** Ridha bin Haji Asmat
Mohd	Mohammad	Pengiran Lela Utama Pengiran Haji ***Mohd.*** Said
Mr	Mister	***Mr.*** Stanley Loh
Muhd	Muhammad	***Muhd.*** Shah Reza Zuzunnurdaraina bin Mohd. Zurimi
No	Number	Pemeriksaan di lokasi terakhir, di bilik sewa ***No.*** 111
Sdn	Sendirian	Majlis penandatanganan kontrak bagi para penerima Biasiswa Brunei Gas Carrier ***Sdn.*** Bhd. (BGC)
Sr	Senior	***Sr.*** Dayang Hajah Norhayati binti Haji Mohd. Yaakub
Vol	Volume	Majlis Pelancaran Kempen Brunei Unified ***Vol.*** 2 yang bertemakan "Mata Hati"

(Note: Abbreviations are Bolded and italic in the example).

Next, unnecessary sentences such as sentences containing long English speeches or a mix of Malay and English words (where it is primarily English) from quoted texts are removed from the corpus. Fig. 2 shows a sample of the removed data. These data are filtered out because the sentence is not entirely in Malay language and contains long English phrases. This might be a challenge to include the sentences in the corpus as we have not addressed the issue of having sentences containing multiple languages and how it may impact on the learning. A total of 12 sentences are discarded. After performing data cleaning, the size of BruCorp is reduced to 3,040 sentences.

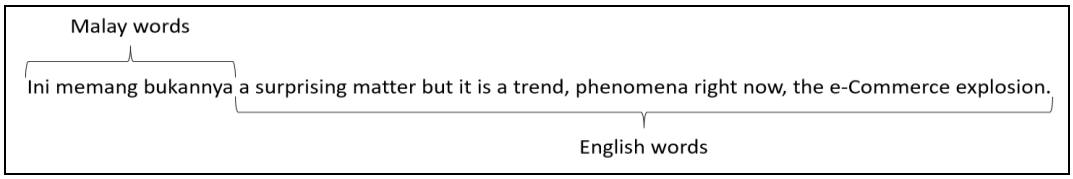

Figure 2. A sample of removed data from the corpus containing a mix of Malay and English words.

D. Data Annotation

Three local language experts manually tag the corpus to represent the gold standard for evaluating NLP systems. The annotation process consists of manual labelling and label verification.

TABLE IV. STATISTICS OF ANNOTATED TOKENS

Tag	Description (English)	No. of tokens	No. of tokens (%)
AUX	Auxiliary	1361	1.18
CC	Conjunction	11,775	10.25
CD	Cardinal number	4934	4.30
DET	Determiner	1452	1.26
IN	Preposition	7268	6.33
JJ	Adjective	2837	2.47
NEG	Negation	475	0.41
NN	Noun	29,254	25.47
NNP	Proper Noun	22,400	19.50
PRP	Pronoun	941	0.82
PT	Punctuation	15,121	13.16
RB	Adverb	1907	1.66
SYM	Symbol	258	0.22
VB	Verb	13,547	11.79
WP	Interrogative (WH Questions)	408	0.36
X	Others	108	0.09
X(KP)	Others (Borrowed Words)	829	0.72

Manual labelling is the first step in the annotation process. The annotators label each word in every sentence based on its definition and context. Some issues were brought up during the annotation process that may cause tag ambiguity for some Malay words. Therefore, the tagset was revised for these special cases in the Bruneian corpus such as the tagging for words with combinations of letters and digits (e.g., "2021M" and "1440H"), and commonly used foreign words (e.g., "Da'wah" and "O-levels"). The corpus was retagged using the revised tagset.

The second step is to verify the tags assigned to the tokens. Each expert manually evaluates and corrects all tagged tokens. This step involves spell-checking and misspelling of tag names. The evaluated corpus is the final output of this process. There is a total of 114,875 annotated tokens and Table IV below shows the distribution of the tags in the Bruneian corpus. The top three tags present are "Noun", "Proper Noun" and "Punctuation".

IV. EXPERIMENT SETUPS

The Bruneian corpus is split into training and testing sets. The training data consists of news articles from BruDirect website [28]. There are two sets of testing data where the first test set is taken from the same news website whereas the second test set is taken from the police government website [29]. In the experiments, we trained the models on first 80% of the news dataset and evaluate the model on two test sets (remaining 20% news articles and all crime articles). Table V below shows how we divide the corpus into two sets.

TABLE V. DATA SPLIT FOR TRAINING AND TESTING SETS

Set	No. of sentences
Train	2026
Test (News Articles)	507
Test (Crime Articles)	507

A. Conditional Random Field (CRF)

CRFs use feature functions that give information about patterns in a sequence. These features are often hand-crafted and can be independent of one another. The model then calculates the tag's probability to be predicted based on the features. We use sklearn-crfsuite to implement CRF experiments with L1 and L2 regularization values set to 0.1. We conducted five sets of experiments using different features combinations and summarized them in Table VI.

- N-grams are sequences of n-sized words. For instance, when n = 1, it is a Unigram, and it takes a single word (e.g., "computer"), whereas when n = 2, it is a Bigram and takes two words (e.g., "computer science").
- The Malay language uses four types of affixations: prefixes, suffixes, circumfixes, and infixes [30]. An affixed word is the result of when a root word (or simply, the base) is combined with any affixation type. Prefixes usually add two or three characters at the beginning of the base. Suffixes usually add two or three characters at the end of the base. Circumfixes are the combination of both prefixes and suffixes to form a single morphological unit. Infixes add an infix to the first consonant of the root word. In the experiments, we have excluded the use of infixes because it is ineffective in the Malay language and to avoid POS tag ambiguity [7, 30].
- Numerical features return a numerical value corresponding to the feature, such as "Word Length" returns the length of the word in that sequence.
- Binary features consist of features provided by other works [14, 29, 32, 33]. These features return a "True" or "False" value, such as if the word appears at the beginning or at the end of the sentence.

TABLE VI. FIVE SETS OF CRF EXPERIMENTS

Set #	Types of Features	Features
1	Baseline (BL)	Unigrams + Bigrams with a context window of 5
2	Affixation	BL + Prefixes + Suffixes
3	Numerical	BL + Ratio + Vowel + Word Length + Sentence Length
4	Binary	BL + Duplicates, First word, Last word, Capitalized, All upper case, All lower case, Hyphen, Has upper case
5	All	BL + Affixation + Numerical + Binary

B. Bi-directional Long Short-Term Memory (BLSTM)

Bi-directional LSTM has the same architecture as LSTM but in bi-directional ways (forward and backward). Our experiments use character and three pre-trained word embedding models, namely Google's Word2Vec, Malay fastText, and Indonesian fastText [34], with the BLSTM model. The experiment's hyperparameter settings are set to word embedding size of 100, character embedding size of 30, hidden layers of 100, dropout value of 0.3, and L2 regularization of 0.001. We used the optimizer Adam with the default value of 0.001 and ran on 50 epochs in batches of 128. We grouped the experiments into three groups: BLSTM only, BLSTM with CRF, and BLSTM without CRF. We also conducted the same experiments with an additional Batch Normalization layer [35] and compared its performance. Table VII summarizes the experiments.

TABLE VII. THREE GROUPS OF BLSTM EXPERIMENTS. W2V MEANS WORD2VEC, FT MEANS FAST TEXT

Group #	Model #	Combinations
1	1	BLSTM
2	2.1	BLSTM + CRF
	2.2	BLSTM + CRF + W2V
	2.3	BLSTM + CRF + CHAR + W2V
	2.4	BLSTM + CRF + FT (Malay)
	2.5	BLSTM + CRF + CHAR + FT (Malay)
	2.6	BLSTM + CRF + FT (Indonesian)
	2.7	BLSTM + CRF + CHAR + FT (Indonesian)
3	3.1	BLSTM + CHAR
	3.2	BLSTM + CHAR + W2V
	3.3	BLSTM + CHAR + FT (Malay)
	3.4	BLSTM + CHAR + FT (Indonesian)

V. RESULTS AND DISCUSSION

A. Main Results

We evaluated the tagging models on two test sets (news articles and crime articles). F-Measure is used for the evaluation metric as it considers both precision and recall of predicted tags; therefore, can measure the model's performance better. Tables VIII–X report the F-Measures of CRF models, BLSTM models without Batch Normalization, and BLSTM models with Batch Normalization, respectively. For CRF experiments, the model with Set 5 features (All) performed the best on both news and crime articles. For BLSTM experiments,

models with a Batch Normalization setting generally performs better with at least 88% of F-Measure for news articles, and at least 82% of F-Measure for crime articles.

Without Batch Normalization set, the F-Measures are at least 82% and 80% for news and crime articles, respectively. When Batch Normalization was not set, BLSTM + CHAR achieved the highest performance with F-Measure of 88.57% on news articles and 85.93% on crime articles. With Batch Normalization set, two models from Group 2, BLSTM + CRF + Malay fastText and BLSTM + CRF + Indonesian fastText, performed the best on news articles with an F-Measure of 89.93%. Meanwhile, BLSTM combined with CRF, CHAR, and Word2Vec word embeddings performed the best with F-Measure of 86.58% on crime articles. However, all BLSTM models and combinations performed worse than CRF models overall.

TABLE VIII. THE PERFORMANCE OF CRF MODELS ON FIVE SETS

Set No.	F-Measure (%) on News Articles	F-Measure (%) on Crime Articles
1	90.69	86.75
2	92.00	89.75
3	90.92	87.16
4	91.31	89.36
5	**92.06**	**90.71**

TABLE IX. THE PERFORMANCE OF DIFFERENT BLSTM GROUPS WITHOUT BATCH NORMALIZATION

Group No.	Model No.	F-Measure (%) on News Articles	F-Measure (%) on Crime Articles
1	1	82.72	80.51
2	2.1	85.34	86.73
	2.2	83.52	80.72
	2.3	86.63	83.26
	2.4	85.37	82.20
	2.5	87.07	83.36
	2.6	84.23	81.97
	2.7	86.73	84.33
3	3.1	**88.57**	**85.93**
	3.2	88.06	85.15
	3.3	86.84	84.63
	3.4	86.52	84.34

TABLE X. THE PERFORMANCE OF DIFFERENT BLSTM GROUPS WITH BATCH NORMALIZATION

Group No.	Model No.	F-Measure (%) on News Articles	F-Measure (%) on Crime Articles
1	1	89.85	86.28
2	2.1	89.90	86.56
	2.2	89.90	85.42
	2.3	89.90	**86.58**
	2.4	**89.93**	84.73
	2.5	89.15	85.49
	2.6	**89.93**	86.36
	2.7	89.43	85.24
3	3.1	89.13	85.79
	3.2	89.31	84.68
	3.3	88.14	82.40
	3.4	89.89	85.74

B. Discussions

Based on the experiments, the results showed that CRF models and BLSTM could predict Standard Brunei Malay; CRF model all features (Unigrams, Bigrams, Affixations, Numeric and Binary) achieved F-Measures of 92.06% and 90.71% on news and crime articles respectively. Prefixes and suffixes are features that provide helpful information to the training models; thus, models with Set 2 and Set 5 could learn the patterns and predict the Malay texts very well and the differences is small. This work is similar to Kurniawan and Aji [8], who also implemented CRF for Indonesian POS tagging and also used affixation information, achieving high accuracy of tagging performance.

Using the Batch Normalization technique greatly improved the performance of most of the BLSTM models. This normalization method also improved performance in other works [36, 37]. This work demonstrated that a batch normalization layer could be added to the BLSTM model architecture for POS tagging tasks. In our BLSTM experiments, the highest F-Measure was achieved by the combining BLSTM, CRF and fastText (Malay and Indonesian) where it scored 89.93% on news articles and 86.58% on crime articles by combining BLSTM, CRF, character embedding and Word2Vec.

Although it was predicted that Malay fastText would improve the tagging the most due to the language's nature, it is not the case. The results also showed that Word2Vec and Indonesian fastText were equally efficient when compared. One possible reason is that both Word2Vec and fastText (Indonesian) models are trained on a large number of words compared to Malay fastText. Google's Word2Vec is trained on 100 billion words and on many languages. Therefore, we assumed a significant lack of tokens to train for Malay fastText. The second reason is that the hyperparameter settings may not be suitable for the Bruneian corpus and can still be experimented with using different values or settings to improve tagging performance. Considering that CRF requires hand-crafted features and BLSTM does not, this proved that BLSTM models could extract the information automatically without human intervention.

Overall, it was observed that CRFs performed slightly better than BLSTMs. CRF achieved the highest F-Measure of 92.06 and 90.71 for news and crime articles, respectively. Other works report similar results where some CRF models obtained higher F-Measures than deep learning models for sequence labeling tasks [38–40]. Deep Learning approaches are known to be more data-demanding than the statistical approach CRF [40] which means BLSTMs require more data to make correct predictions. Therefore, it results in multiple incorrect predictions by BLSTM models where the least frequent tags are not handled well due to the imbalance of tags which lowers the overall F-Measure. In contrast, CRFs are able to predict most tags correctly, including the uncommon tags found in the training set, such as "NEG" and "SYM", thus giving higher F-Measure. Furthermore, CRF takes context into account when predicting the output; hence more tokens are accurately predicted than BLSTM. Context is a valuable and crucial component in NLP that can help build accurate NLP models [41–43], specifically for POS tagging and Named-Entity Recognition (NER).

VI. Conclusion

In this paper, we have described the development of a Bruneian corpus which consists of over 3000 sentences. In the process of corpus development, we also design a Malay POS tagset by extracting common tags from existing Malay POS tagging works and adding a few additional tags to solve POS tag ambiguity. We also introduced abbreviations commonly found in the Bruneian texts and included the abbreviations in the sentence tokenizer so it can separate Malay sentences more accurately. Furthermore, we implemented CRF and BLSTM for Bruneian POS tagging. We compared the tagging performance of the different feature sets of CRF and combinations of BLSTM models with CRF, character embeddings, and pre-trained word embedding models. The experimental results showed that CRF models outperform the BLSTM model combinations. The lack of training data for the least frequent tags is one of the potential factors that contribute to the poor performance of BLSTM models.

For future works, BLSTM models can be improved by experimenting with different hyperparameter settings and increasing the amount of data corpus. Other techniques can be explored to combine with the BLSTM models, such as CNN and BERT, to improve the POS tagging performance further.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Izzati Mohaimin conducted the research, collected, and analyzed the data and wrote the initial draft and final version of the paper. Rosyzie A. Apong supervised the work, validated the analysis results, edited, and reviewed the overall structure of the paper and project. Ashrol R. Damit supervised the work and provided the necessary information for linguistic resources including the tagging and validation of corpus. All authors had approved the final version.

Funding

This project is funded and supported by the Ministry of Education (MOE) of Brunei Darussalam, and Universiti Brunei Darussalam.

References

[1] T. Baldwin and S. Awab, "Open source corpus analysis tools for Malay," in *Proc. the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, 2006.

[2] J. T. Collins and A. Zaharani, "The Malay language and ethnic identity in modern Malaysia," *Akademika*, vol. 55, pp. 133–148, 1999.

[3] A. Omar, "The Malay language in Malaysia and Indonesia: From lingua franca to national language," *Asianists' ASIA*, vol. 2, pp. 1–21, 2001.

[4] N. Phillips, "Differences between Bahasa Indonesia and Bahasa Malaysia," *Indonesia Circle. School of Oriental & African Studies. Newsletter (Currently known as Indonesia and the Malay World)*, vol. 1, no. 2, pp. 7–9, 1973, doi: 10.1080/03062847308723526

[5] A. Clynes, "Brunei Malay: An overview," *Occasional Papers in Language Studies*, vol. 7, pp. 11–43, 2001.

[6] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, September 2011, doi: 10.1136/amiajnl-2011-000464

[7] R. Alfred, A. Mujat, and J. H. Obit, "A Ruled-Based Part of Speech (RPOS) tagger for Malay text articles," in *Proc. the Asian Conference on Intelligent Information and Database Systems*, 2013, pp. 50–59, doi: 10.1007/978-3-642-36543-0_6

[8] K. Kurniawan and A. F. Aji, "Toward a standardized and more accurate Indonesian part-of-speech tagging," in *Proc. the 2018 International Conference on Asian Language Processing, IALP*, Bandung, 2018, pp. 303–307, doi: 10.1109/IALP.2018.8629236

[9] E. Brill, "A simple rule-based part of speech tagger," in *Proc. the Third Conference on Applied Natural Language Processing (ANLC'92)*, Trento, 1992, pp. 152–155, doi: 10.3115/974499.974526

[10] T. Brants, "TnT—A statistical part-of-speech tagger," in *Proc. the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, 2000, pp. 224–231, doi: 10.3115/974147.974178

[11] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. the Eighteenth International Conference on Machine Learning*, San Francisco, 2001, pp. 282–289.

[12] K. Darwish, *et al.*, "Multi-dialect Arabic POS tagging: A CRF approach," in *Proc. the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 2018, pp. 93–98.

[13] W. AlKhwiter and N. Al-Twairesh, "Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM," *Computer Speech and Language*, vol. 65, 2021, doi: 10.1016/j.csl.2020.101138

[14] F. Nooralahzadeh, C. Brun, and C. Roux, "Part of speech tagging for French social media data," in *Proc. COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, 2014, pp. 1764–1772.

[15] S. Fu, N. Lin, G. Zhu, and S. Jiang, "Towards Indonesian part-of-speech tagging: Corpus and models," in *Proc. LREC 2018 Workshop on Belt and Road LRE*, Paris, 2018, pp. 2–7.

[16] S. Briandoko, A. R. Dewi, and M. A. Setiawan, "Comparison of the conditional random field and hidden Markov model algorithm in the Indonesian tagging post," *Engineering: Information Technology, Computer Science and Management*, vol. 2, pp. 23–27, 2018. (in Indonesian)

[17] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic part of speech tagging for Bahasa Indonesia," in *Proc. the 3rd International MALINDO Workshop, Colocated Event ACL-IJCNLP*, 2009.

[18] T. T. Wai, "Myanmar language part-of-speech tagging using deep learning models," *International Journal of Scientific and Engineering Research*, vol. 10, no. 3, 2019.

[19] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint, arXiv:1508.01991, 2015.

[20] H. Tang, H. Hammarström, and Y. Shao, "Bidirectional LSTM-CNNs-CRF models for POS tagging," M.S. thesis, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, 2018.

[21] P. T. Tan, *et al.*, "Evaluating LSTM networks, HMM and WFST in Malay part-of-speech tagging," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, pp. 79–83, 2017.

[22] H. Mohamed, N. Omar, and M. J. A. Aziz, "Statistical Malay Part-of-Speech (POS) tagger using hidden Markov approach," in *Proc. the 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011*, 2011, pp. 231–236, doi: 10.1109/STAIR.2011.5995794

[23] N. A. Halid and N. Omar, "Malay part of speech tagging using ruled-based approach," *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 6, no. 2, pp. 91–107, 2017.

[24] M. P. Hamzah and S. F. N. S. Kamaruddin, "Part of speech tagger for Malay language based on words morphology," in *Proc. the International Symposium on Research in Innovation and Sustainability 2014 (ISoRIS'14)*, 2014, pp. 1499–1502.

[25] B. C. M. Xian, *et al.*, "Benchmarking Mi-POS: Malay part-of-speech tagger," *International Journal of Knowledge Engineering*, vol. 2, no. 3, pp. 115–121, 2016, doi: 10.18178/ijke.2016.2.3.064

[26] N. Zamin, A. Oxley, Z. Abu Bakar, and S. A. Farhan, "A lazy man's way to part-of-speech tagging," in *Proc. the Knowledge Management and Acquisition for Intelligent Systems, Lecture Notes in Computer Science*, 2012, vol. 7457, pp. 106–117, doi: 10.1007/978-3-642-32541-0_9
[27] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," *in Proc. the International Conference on Asian Language Processing 2014, IALP 2014*, 2014, pp. 66–69, doi: 10.1109/IALP.2014.6973519
[28] BruDirect Website. (2022). [Online]. Available: https://www.brudirect.com/
[29] Pasukan Polis Diraja Brunei Website. (2022). [Online]. Available: https://www.polis.gov.bn/Theme/Home.aspx
[30] B. Ranaivo-Malançon. (2004). Computational analysis of affixed words in Malay language. [Online]. Available: https://www.researchgate.net/publication/254350731_COMPUTATIONAL_ANALYSIS_OF_AFFIXED_WORDS_IN_MALAY_LANGUAGE
[31] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. HLT-NAACL 2003*, 2003, pp. 173–180.
[32] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proc. the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000, pp. 63–70.
[33] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
[34] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," arXiv preprint, arXiv: 1802.06893, 2018.
[35] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint, arXiv: 1502.03167, 2015.
[36] W. Li, G. Cheng, F. Ge, P. Zhang, and Y. Yan, "Investigation on the combination of batch normalization and dropout in BLSTM-based acoustic modeling for ASR," in *Proc. the Annual Conference of the International Speech Communication Association*, 2018, pp. 2888–2892, doi: 10.21437/Interspeech. 2018-1597
[37] P. Zelasko, *et al*., "Punctuation prediction model for conversational speech," arXiv preprint, arXiv: 1807.00543, 2018.
[38] J. Gugglberger. (2020). Comparing CRF and BI-LSTM networks for Named Entity Recognition (NER). PhD thesis, Institute of Computer Science, University of Innsbruck, Innsbruck, Austria. [Online]. Available: https://github.com/moejoe95/crf-vs-rnn-ner
[39] R. D. Deshmukh, "Comparison of generative and discriminative models of part of speech taggers for Marathi language," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 16–21, 2018, doi: 10.26438/ijcse/v6i10.1621
[40] J. Rao, F. Ture, and J. Lin, "Multi-task learning with neural networks for voice query understanding on an entertainment platform," in *Proc. the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 636–645, doi: 10.1145/3219819.3219870
[41] S. Warjri, P. Pakray, S. A. Lyngdoh, and A. K. Maji, "Part-of-Speech (POS) tagging using Conditional Random Field (CRF) model for Khasi corpora," *International Journal of Speech Technology*, vol. 24, pp. 853–864, 2021, doi: 10.1007/s10772-021-09860-w
[42] D. U. Patton, *et al*., "Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, New York, 2020, pp. 337–342, doi: 10.1145/3375627.3375841
[43] M. C. Elish and D. Boyd, "Situating methods in the magic of big data and artificial intelligence," *Communication Monographs*, vol. 85, no. 1, pp. 57–80, 2018, doi: 10.1080/03637751.2017.1375130