Part-of-Speech Tagger for Malay Social Media Texts

Siti Noor Allia Noor Ariffin
sitinoorallia@gmail.com
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

Sabrina Tiun
sabrinatiun@ukm.edu.my
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

ABSTRACT

Processing the meaning of words in social media texts, such as tweets, is challenging in natural language processing. Malay tweets are no exception because they demonstrate distinct linguistic phenomena, such as the use of dialects from each state in Malaysia; borrowing foreign language terms in the context of Malay language; and using mixed languages, abbreviations and spelling errors or mistakes in sentence structure. Tagging the word class of tweets is an arduous task because tweets are characterised by their distinctive style, linguistic sounds and errors. Currently, existing works on Malay part-of-speech (POS) are based only on standard Malay and formal texts and are thus unsuitable for tagging tweet texts. Thus, a POS model of tweet tagging for non-standardised Malay language must be developed. This study aims to design and implement a non-standardised Malay POS model for tweets and performs assessment on the basis of the word tagging accuracy of test data of unnormalised and normalised tweet texts. A solution that adopts a probabilistic POS tagging called QTAG is proposed. Results show that the Malay QTAG achieves best average POS tagging accuracies of 90% and 88.8% for normalised and unnormalised test datasets, respectively.

Keywords: part-of-speech; informal Malay text; Malay POS tagger; Malay tweet; QTAG

INTRODUCTION

Part-of-speech (POS) is a category that divides words on the basis of their use and functions in a sentence. For example, the English language has eight major POS: nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections. In the meantime, Malay language has 18 major POS listed by Dewan Bahasa Dan Pustaka (DBP), a government organisation responsible for managing the practice of the Malay language and literature in Malaysia (Hassan, 1986). Gimpel et al. (2011) and Antony, Mohan and Soman (2010) stated that POS tagging plays an important role in the linguistic pipeline and is a basic form of syntactic analysis that has numerous applications in natural language processing (NLP), such as sentiment analysis (Altawaier & Sabrina, 2016; Nielsen, 2011) and named entity recognition (Alshaikhdeeb & Ahmad, 2016). Chowdhury (2003) highlighted that NLP is an area of research and application that deals with the ability of a computer programme to understand and process human language in large amounts of natural language data. A challenge in NLP is to process the meaning of words in social media texts, such as tweets, because tweets are written freely without maintaining a formal grammar and correct spelling and often use abbreviated words (Java, Song, Finin & Tseng, 2007). Figure 1 shows an example of a tweet that demonstrates an NLP challenge.

eISSN: 2550-2131
ISSN: 1675-8021

```
Aq pelik viral tambang bas merah 4 genih. Yg viral tu pulak
budak2 yg tgh bercuti kat malaysia. Geng, aq dlm bas skg ni.
Harga masih 3 genih. Bas mini tu je 4 genih setakat ni. xkn nk
aq selfi dlm bas ni pulak bru caye
```

FIGURE 1. Example tweet that is challenging for Malay POS tagging

POS tagger was originally developed by Toutanova and Manning (2000) for the English language and obtained an accuracy of 96.86% on the Penn Treebank (PTB), a parsed text corpus that annotates syntactic or semantic sentence structures. Later, the tagger was improved in terms of speed, performance, usability and support for other languages (Toutanova, Klein, Manning & Singer, 2003), thereby resulting in an accuracy of 97.24% and error reduction of 4.4% on Wall Street Journal articles PTB.

In this study, we construct a Malay POS tagger designed especially for data from Twitter, a popular microblogging service (Gimple et al., 2011). In contrast to current Malay POS tagger studies (Anbananthen et al., 2017; Chekima & Rayner, 2017; Xian et al., 2016), this study designs a Malay POS tagger exclusively for the informal Malay language that is composed of dialects (regional variation), grammatical and typographical errors and abbreviations, especially in social media texts (tweets). The contributions of this study are listed as follows:

- A POS tag set for Twitter is developed
- A total of 500 tweets can be manually tagged
- QTAG (Tufis & Mason, 1998; Mason & Tufis, 1997) features are used for Twitter POS tagging and evaluated using experiments
- The proposed annotated corpus and trained POS tagger can be utilised in the NLP research field and educational society

The Malay POS tagger, which was created using QTAG (Tufis & Mason, 1998; Mason & Tufis, 1997) models, is a supervised machine learning (ML) POS tagging approach that requires a large amount of annotated training corpus data to tag the identified data accurately. Elworthy (1995) argued that the size of the POS tagger set affects the performance of the tagging. An appropriate training corpus has a total fraction between 100,000 and more than one million words in the same corpus. Although some tagging has been programmed to learn language models from raw texts (without annotations), the taggers still require verification after the output is released, and bootstrapping procedures are needed to ensure that the tagging reaches the minimum level of error rate. A large tagger set indicates a large size of the required training corpus (Berger, Della Pietra & Della Pietra, 1996). A POS tagger for informal Malay social media texts must be developed because of its high demand in industries, such as NLP-related research fields and educations.

The informal Malay language is formed from the modification of words in the Malay Standard or is derived from other languages. This type of language is often used especially in urban communities and may be difficult to understand amongst the previous generation (Maslida, 2018); example words include *usha* (*perhati*)/observe, *skodeng* (*intai*)/peek, *cun* (*lawa*)/pretty and *poyo*, *slenge* (*buruk*)/bad. New pronouns are also designed using existing prefix combinations with words that refer to people, such as *kitorang* (*kita*/us + *orang*/people, replacing the word *kami*/we), *korang* (*kau*/you + *orang*/people, to refer to many people and replace the word *kalian*/all of you) and *diaorang* or *dorang* (*dia*/them + *orang*/people, substitute for the word *mereka*/they). Communicating via social media has encouraged the switch coding of Malay Standard and dialect, such as the Javanese dialect as mentioned by Karim and Maslida (2015).

Nurul et al. (2015) mentioned that the change of one language to another within the same utterance or in the same oral/written text is referred to as code-mixing. By contrast, Muysken (2000) defined lexical items and grammatical features of two languages found in the same sentence as code-mixing. Li (1998) explained that any admixture of linguistic elements of two or more language systems in the same utterance at various levels (i.e. phonological, lexical, grammatical and orthographical) indicates a code-mixing. Nurul et al. (2015) also stated that, in a multilingual society, such as in Malaysia, code-mixing is a common phenomenon that generates mixed languages. Yang et al. (2016) also acknowledged that a mix of different cultures often creates a mix of languages in a sentence, particularly from informal sources. Examples of loan terms used are Best*lah tempat ni* (This is a good place) and *Kau ni* terror*lah* (You're awesome). Table 1 displays some short forms of words that are commonly used by Malaysian teenagers.

TABLE 1. List of abbreviations often used by Malaysian teenagers

No. of Abb.	Types of Abbreviation	Normal Phrase	Example	Meaning
1	Replace tidak with the letter x	Tidak boleh	X boleh	Cannot
2	After reduplication	Hari-hari	Hari2	Everyday
3	Eradicate vowel letters	Bangun	Bgn	Get up
4	Eradicate the letter r	Terserang	Terseang	Attacked
5	Eradicate affix	Kekasih	Kasih	Sweetheart
6	Eradicate initial letter	Itu	Tu	That
7	Eradicate last letter	Tidur	Tidu	Sleep
8	Combine words	Macam mana	Camne	How

Source: Nasiroh, Ahmad, Nur and Siti (2017)

POS TAGGING OF SOCIAL MEDIA TEXTS IN OTHER LANGUAGES

Anbananthen et al. (2017) and Xian et al. (2016) conducted POS tagging of Malay social media texts. However, none of them focused on developing a tag set and a corpus especially for the informal Malay language that comprises dialects, grammatical and typographical errors and abbreviations. POS tagging of social media texts in other languages has been widely explored. Several novel related works on this POS tagging along with the results (the percentage of accuracy) are discussed below.

Owoputi et al. (2013) used POS tagging to improve English POS tagging accuracy for online conversational texts (e.g. those used in Twitter and Internet Relay Chat) by evaluating the use of large-scale unsupervised word clustering and new lexical features with a first-order maximum entropy Markov model tagger, which resulted in a tagging accuracy of 93%. Derczynski, Ritter, Clark and Bontcheva (2013) performed detailed error analysis of existing taggers and identified and evaluated techniques to improve the performance of English POS tagging. They achieved a POS tagging accuracy of 88.7% and reduction rates of 26.8% and 12.2% for token and sentence errors, respectively.

Nooralahzadeh, Brun and Roux (2014) constructed a French POS tagger for social media data, such as data from Twitter, Facebook and forums, using a linear-chain CRF model that is enriched with abundant morphological, orthographic, lexical and large-scale word clustering features. They successfully obtained a high POS tagging accuracy of 91.9%.

Albogamy and Ramsay (2015) evaluated and performed a detailed error analysis of the Arabic POS tagger. They found that the Arabic POS tagger performance was excellent on Modern Standard Arabic texts with a POS tagging accuracy of 96%–97% compared with the performance on the Arabic tweets with a POS tagging accuracy of 46%–65%. After some improvement, the Arabic POS tagger achieved a POS tagging accuracy of 79%.

Gui et al. (2017) incorporated large-scale unlabelled in-domain and labelled out-of-domain and in-domain data for the Twitter POS tagging task. They used target preserved adversarial neural network to learn domain-invariant representations through in-domain and out-of-domain data and constructed a cross-domain POS tagger through the learned representations. Gui et al. (2017) then obtained a similar tagging accuracy result to that of Owoputi et al. (2013).

Abdulkareem and Sabrina (2017) proposed designing and implementing speech tagging models for Arabic tweets through investigating various models of ML, such as K-nearest neighbour, Naive Bayes and decision tree. They then produced an automatic feature-rich POS tagger and conducted tweet analysis using an ML classifier. The results showed a POS tagging accuracy of 87.97%.

In the meantime, van der Goot, Plank and Nissim (2017) studied the impact of normalisation on POS tagging in a realistic setup by comparing normalisation of unknown words with fully automatic normalisation model. They then evaluated the normalised corpus using the word embedding and self-training approach. The word embedding technique obtained a POS tagging accuracy of 90%.

Amongst all the studies on POS tagging reviewed, the highest POS tagging average is 93% for English language which was achieved by Owuputi et al. (2013). These authors mostly used the Standard English POS tag set called PTB. The reviewed studies show that an effective approach for POS tagging of social media texts is the supervised ML approach, and the most studied languages for this task are English, French and Arabic.

POS TAGGING OF MALAY SOCIAL MEDIA TEXTS

As mentioned in the previous section, POS tagging of Malay social media texts has been recently conducted by Anbananthen et al. (2017) and Xian et al. (2016). Nevertheless, none of them developed a Malay tag set and a corpus solely for the informal Malay language that comprises dialects, grammatical and typographical errors and abbreviations, especially in the Twitter domain. Several novel related works on Malay POS tagging of social media texts and their results (the percentage of accuracy) are discussed below.

Chekima and Rayner (2017) conducted sentiment analysis of informal Malay social media texts by using a framework to handle common challenges posed by these texts. They discussed features on managing Bahasa Rojak (mix-code language), Bahasa SMS, emoticons and valance shifter. Thereafter, they designed a RojakLex lexicon consisting of four different lexicons combined together: MySentiDic (a Malay lexicon), English lexicon (translated version of MySentiDic), emoticon lexicon (combination of nine different renowned used online emoticons) and neologism lexicon (consists of neologism words commonly used in Malay social media texts). The proposed framework successfully achieved an accuracy of 79.28%.

Anbananthen et al. (2017) compared stochastic and rule-based POS tagging approaches to deal with ambiguous and unknown words for Malay online texts. The results showed no significant difference between the average accuracy of rule-based (93.4%) and stochastic (92.1%) approaches for ambiguous word tagging. However, for unknown word tagging, the average accuracy obtained by rule-based was higher with 89.9% than that of stochastic tagger with 85.6%. The overall average accuracy of the rule-based approach was 92.9%, whereas that of the stochastic tagger was 91.4%. Xian et al. (2016) developed a benchmarking Mi-POS of Malay POS tagger using a probabilistic approach with context information. They compared their probabilistic Malay POS (Mi-POS) against the rule-based and HMM approaches. The results showed that Mi-POS outperformed other Malay POS

tagger approaches with accuracies of 95.16% and 81.12% for tagging new words from the same training corpus and words from different corpora types, respectively.

The reviewed studies indicate the possible directions for Malay POS tagging of social media texts: (i) a tagged corpus with suitable POS tag set must be developed, and (ii) an automatic probabilistic POS tagging based on QTAG for the informal Malay language must be proposed (Tufis & Mason, 1998; Mason & Tufis, 1997). In the subsequent sections, the reasons for selecting QTAG will be justified. A suitable POS tag set and a Malay tweet corpus annotated with POS labels will also be developed.

MALAY POS TAGS

Works on Malay POS tagging were conducted by Rayner, Adam and Joe (2013), Juhaida, Khairuddin, Mohammad and Mohd (2013), Mohamed, Nazlia and Mohd (2015), Halid and Nazlia (2017), Arbak (2005), and Hock (2009). These researchers mentioned that the criteria used for Malay POS tagging are roughly similar to those for Greek and English tagging.

Nevertheless, the English POS tag set is inappropriate for the Malay language, especially for the informal Malay language. For example, the PTB POS tag set contain several POS tags that are inappropriate for the Malay language, such as VBZ (verb, third person singular present): *miss + es = misses*, VBP (verb, non-third person singular present): *run + s = runs*, VBD (verb, past tense): *play + ed = played*, VBN (verb, past participle): *sing–sang–sung* and VBG (verb, gerund or present participle): *go + ing = going*. The reason is that the basis for such POS tags is that the verbs are associated with the subject and the time when the action occurred, whereas the verb in the Malay language is not clearly associated with the execution (whether plural or singular) and the time when the act was performed, such as *rindu*/miss, *membuat*/do, *menyanyi*/sing and *pergi*/go.

The PTB POS tag set is clearly incomplete because of the absence of a POS tag for the collective nouns that exist in the Malay language. Therefore, this study aims to develop a suitable POS tag set that caters especially to informal Malay social media texts prior to normalising, tokening and annotating. A Malay POS tag set depends on the various uses or tendencies that scholars found in various dictionaries, textbooks and/or linguistic computing research. Knowles and Zuraidah (2006) stated that the POS tag set of Dewan Bahasa Dan Pustaka (DBP) is acceptable because DBP is a Malaysian government agency responsible for any issues concerning the Malay language in Malaysia. The POS tag set in DBP is similar to the POS tag set used in other Malay dictionaries, such as those used by Arbak (2005), Gimpel et al. (2011) and Hawkins (2008).

Table 2 shows a comparison of primary DBP POS classes with the Malay POS tag sets used by Hawkins (2008). However, the coverage is larger than that of DBP POS tag sets Therefore, Hawkins (2008)'s 21 POS tag sets are used in this study instead.

TABLE 2. Comparison of DBP POS tag sets against Hawkins (2008)'s POS tag sets

DBP POS Tag Set	Hawkins (2008)'s POS Tag Set	Explanations
N	KN	Kata Nama/*Noun*
K	KK	Kata Kerja/*Verb*
S	ADJ	Adjektif/*Adjective*
I	KSN	Kata Sendi Nama/*Preposition*
–	KB	Kata Bantu/*Auxiliary verb*
–	KG	Kata Ganti/*Pronoun*
H	KH	Kata Hubung/*Conjunction*
A	ADV	Adverba/*Adverb*
–	SR	Kata Seru/*Interjection*
B	KT	Kata Tanya/*Question*
–	KBIL	Kata Bilangan/*Cardinal*

–	KPM	Kata Pemeri/*Narrator*
–	KKT	Kata Keterangan/*Statement*
–	KP	Kata Penguat/*Command*
–	KPB	Kata Pembenar/*Justified word*
W	–	Wacana/*Discourse*
–	IMB	Imbuhan/*Affix*
–	AWL	Awalan/*Prefix*
–	AKH	Akhiran/*Suffix*
–	KEP	Kependekan/*Short form*
#	–	Nombor/*Number*
$	–	Simbol Wang/*Money symbol*
%	–	Simbol Huruf/*Alphabet symbol*
D	–	Dektif/*Deictic*
G	–	Kata Ganti Nama/*Pronoun*
L	–	Senarai/*List*
P	–	Penentu/*Indicator*
X	KNF	Kata Nafi/*Deny*
Z	–	Perkataan Asing/*Foreign word*
–	UNG	Frasa/*Phrase*

Source: Mohamed et al. (2015), who compared DBP POS tag sets with those of Hock (2009)

Klitik (@KG), which is an abbreviation of a noun, is important in the Malay language because it determines the function of the noun, such as *-ku, -kau, -mu* and *-nya*. Another particles (#E) that exist other than *klitik* are *-lah* and *-kah*. The expression (UNG) is for Malay POS tag speech phrases, such as *kud*, *dok* and *ea.* Given that social media texts contain many *klitik* and *particles*, words connected with *klitik* and *particles* must be considered in this study. Figure 2 shows an example of a Malay corpus annotated with POS that contains the *klitik* and/or *particle* terms.

```
… haish_SR memalam_KNG dok_KK sorang2_KNG kat_KSN rmh_KN ni_KN lahh_#E nk_KK dngaq_KK baby_KN
nangis_KK pon_KPB satgi_KKT aq_KGNP yg_KH nangis_KK kott_UNG
… aku_KGDP habaq_KK banyak_ADJ kali_KKT dah_ADJ tiap_KN kali_KKT nak_KK keluaq_KK benda_KGNT
sama_ADJ dok_UNG
```

FIGURE 2. Example of Malay POS corpus

QTAG POS TAGGER

QTAG POS tagging is based on a probabilistic approach. Its basic algorithm is straightforward. Firstly, the tagger searches the dictionary for all possible tags that the current token may have along with their respective lexical probabilities (i.e. the probability distribution of the possible tags for the word form). Then, these probabilities are combined with the contextual probability for each tag to occur in a sequence preceded by the two previous tags. The tag with the highest combined score is selected. Two further processing steps also consider the scores of the tag as the second and first elements of the triplet as the following two tokens are evaluated (Tufis & Mason, 1998; Mason & Tufis, 1997).

The QTAG model developed by Tufis and Mason (1998) works by combining two sources of information: one is a dictionary of words with their possible tags and the corresponding frequencies, and the other is a matrix of tag sequences with associated frequencies. These resources can easily be generated from a pre-tagged corpus (Tufis & Mason, 1998; Mason & Tufis, 1997). Tufis and Mason (1998) stated that tagging works on a window of three tokens with two dummy words at the beginning and end of the text. Tokens are read and added to the window that is shifted by one position to the left each time. The token that 'falls' out of the window is assigned a final tag. The tagging steps by Tufis and Mason (1998) are listed as follows:

1. Read the next token
2. Search it in the dictionary
3. If not found, then guess the possible tags
4. For each possible tag
   a. Calculate P_w = P (tag|token), which is the probability of the token to have the specified tag
   b. Calculate P_c = P (tag|t_1, t_2), which is the probability of the tag to follow tags t_1 and t_2
   c. Calculate $P_{w,c} = P_w * P_c$, which is the joint probability of the individual tag assignment together with the contextual probability
5. Repeat the computation for the two other two tags in the window but using different values for the contextual probability: the probabilities of the tag being surrounded and followed by the two other tags, respectively.

The POS QTAG tagging algorithm can be simplified as shown in Figure 3.

Algorithm: QTAG POS Tagging	
1. Input: word (w)	
2. Output: tokenised and annotated word	
3. Start	
4. Calculate the probability of the token, $P_w = P\ (tag\|token)$	(1)
5. Calculate the probability of the tag, $P_c = P\ (tag\|\ t_1,\ t_2)$	(2)
6. Calculate the joint probability, $P_{w,c} = P_w\ *\ P_c$	(3)

FIGURE 3. Simplified POS QTAG algorithm (Tufis & Mason, 1998)

QTAG POS tagging is a simple and easy approach to adopt to other languages, such as Romanian (Tufis & Mason, 1998), German (Cox, 2010) and Norwegian (Nøklestad & Søfteland, 2007). QTAG has also been a favoured approach for POS tagging of minority languages, such as the Vietnamese language (Nguyen, Vu, & Le-Hong, 2003) and the subgroup of the German language called Plautdietsch (Cox, 2010), due to its simplicity with a minimal number of required resources (a corpus with annotated POS tags and a set of POS tags). The successful and encouraging results (an accuracy in the range of 80%–90%) obtained from these studies using QTAG are the basis of the use of QTAG as the approach to build the Malay POS tagger for Malay social media texts in the current study. The same range of results is expected to be obtained from this study.

Figure 4 shows the steps of this algorithm.

eISSN: 2550-2131
ISSN: 1675-8021

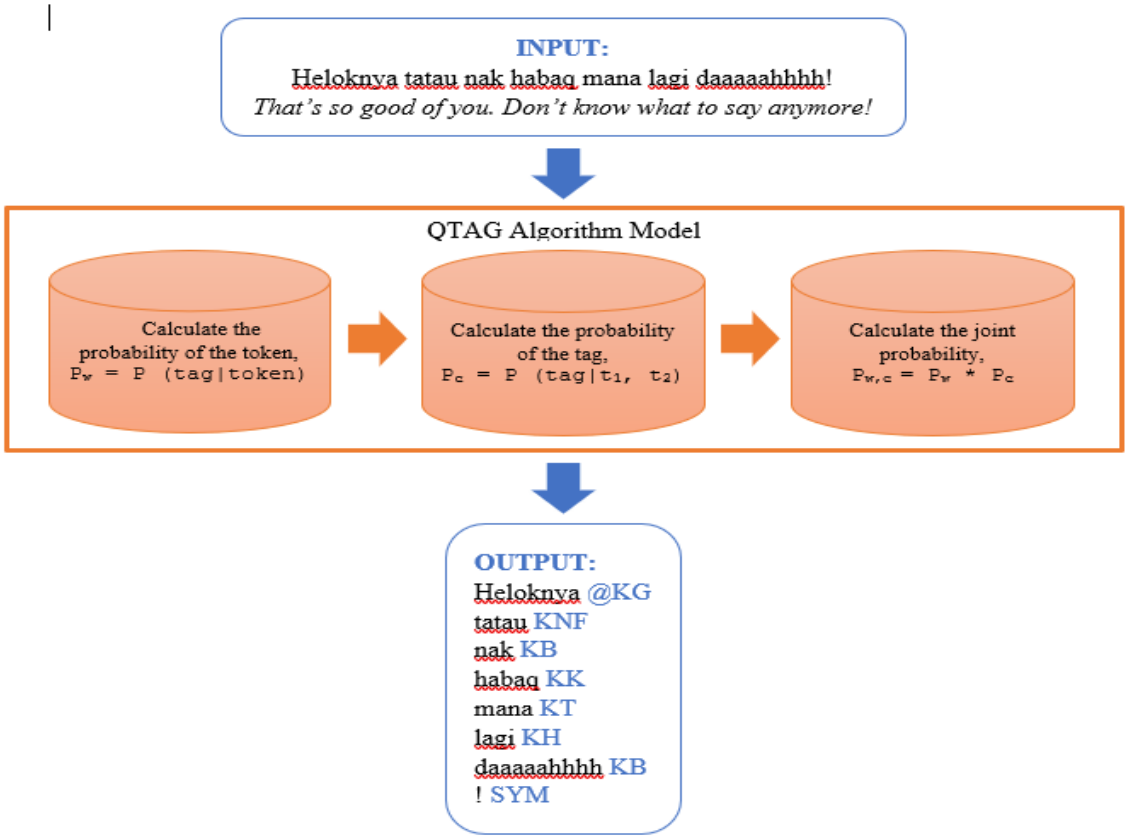

FIGURE 4. Function of the POS QTAG algorithm

METHOD

POS tagger for (informal) Malay social media texts (Twitter) is generally a simple architecture that consists of only two sub-modules: (i) pre-processing sub-modules and (ii) Malay QTAG tagger. The input in the latter sub-module is a raw tweet extracted from twitter.com and generates an output of tokenised and annotated corpus. Figure 5 display the architecture of the proposed Malay POS tagger.

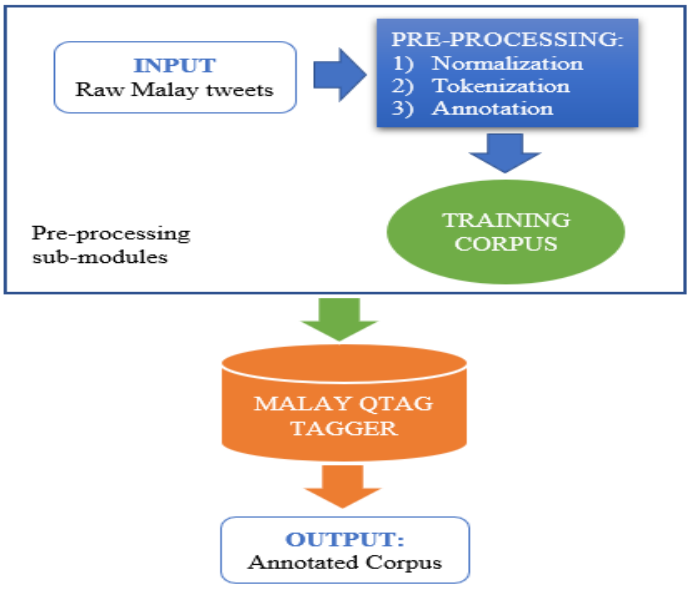

FIGURE 5. Architecture of the Malay QTAG POS tagger.

The development of the Malay POS tagger for social media texts (Twitter) involves two main phases: (i) creating the social media text corpus annotated with POS and (ii) developing the Malay QTAG by adopting the English QTAG developed by Tufis and Mason (1998). In the subsequent subsection, the construction of the Malay POS corpus, as well as the description of the Malay QTAG, will be provided.

MALAY POS CORPUS

The construction of the Malay QTAG tagger requires a set of training corpus, that is, a manually tokenised and annotated corpus. In this corpus, the texts (tweets) are taken (collected) from Malay Twitter accounts (only informal Malay tweets). Tweets that comprise dialects (regional variation), grammatical and typographical errors and abbreviations are used as the training corpus. Statistically, the training corpus contains 70% of informal Malay texts on average, and the rest are standard Malay texts. The following steps are taken to produce a set of training corpus.

Firstly, the texts (tweets) of the corpus are collected manually from twitter.com with only Malay-written tweets accepted. The collected tweets, which are referred to as raw Malay tweets/corpus, undergo the pre-processing phase including normalising, tokenising and annotating. The normalisation phase eliminates any punctuations or symbols in the corpus. The tokenisation phase separates the corpus sentences into tokens on the basis of the root word, spacing between sentences and new lines. The annotating phase manually tags the tokenised corpus with (informal) the pre-designed Malay POS tag set. As a result, a completed annotated corpus labelled as the training corpus is ready to be used as the training data for the Malay QTAG.

Figure 6 illustrates the pre-processing phase, and Table 3 displays the summary of statistical information of the Malay POS tagged (annotated) corpus.

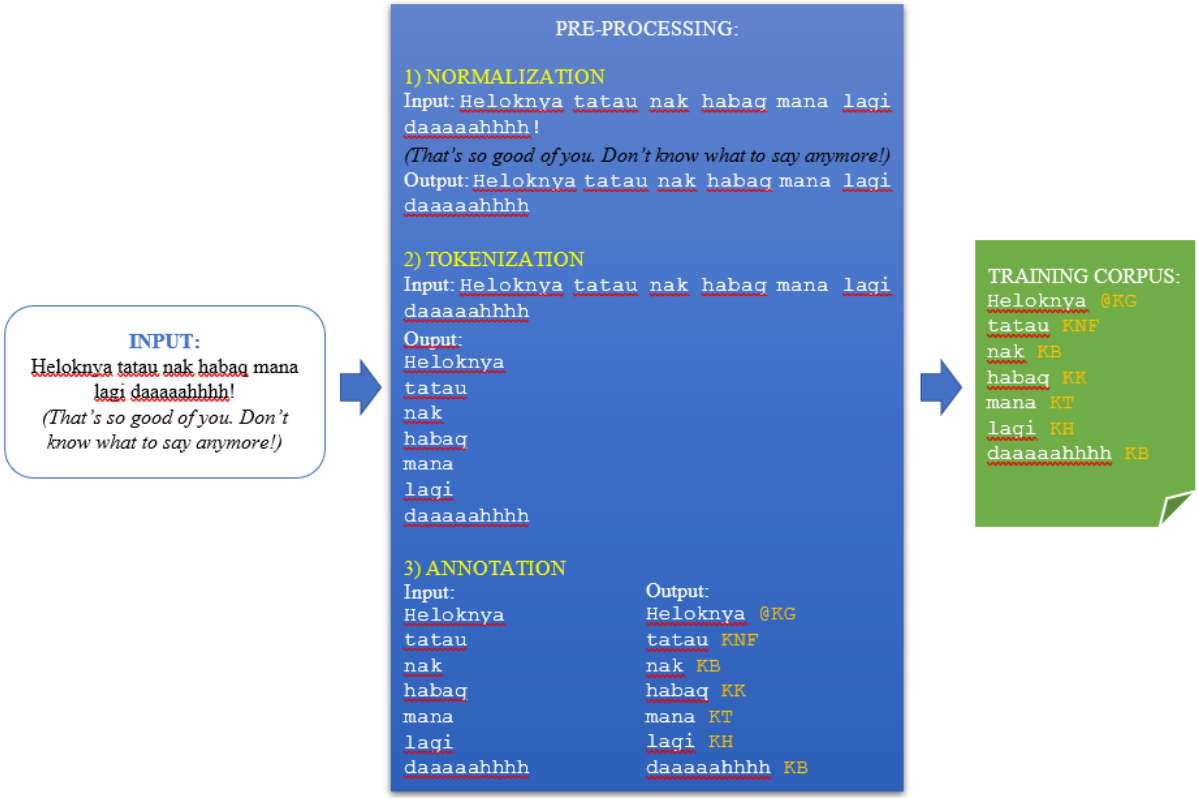

FIGURE 6. Process of the pre-processing sub-modules

TABLE 3. Summary of statistical information of the Malay POS tagged corpus

Items	Value
No. of Tweets	300
No. of Words	5513
No. of Lexical Item	2527
No. of Errors	680
Tag Sets	38

MALAY QTAG

The Malay QTAG POS tagger is designed by adopting the probabilistic QTAG POS tagger created by Tufis and Mason (1998) for the English language. We choose QTAG because its algorithm is straightforward, which eases its use to study POS of any minority languages and various languages, such as English, Thailand and Vietnamese in Cox (2010), Sornlertlamvanich, Charoenporn and Isahara (1997) and Tran, Le, Ha and Le (2009), respectively. However, none of these researchers studied or developed a QTAG POS tagger exclusively for the Malay language.

The design of the Malay QTAG POS tagger begins with the creation of a resource file (`.dat` file). The resource file is built from the training corpus created in the pre-processing sub-modules, which we name `MAZI_pretagged.txt`, and the Malay POS tag set, which we call `MalayTagset.txt`. The following command is then executed to produce the resource file (`MalayTagset.dat`).

```
java -cp qtag.jar qtag.LexiconCreator MalayTagset.dat < MAZI_pretagged.txt
```

The programme that contains the tagger (`QTAG.jar`) and the generator (`qtag.LexiconCreator`) for creating the resource file (`.dat`) can be downloaded from 'Tagging with QTAG' (2007). When the resource file is ready to use, the following command is executed to test the Malay QTAG tagger by inserting a new raw Malay tweet as `input.txt`.

```
java -jar qtag.jar MalayTagset.dat < input.txt > output.txt
```

The evaluation of the new raw Malay tweets (test corpus) is performed by manually checking and correcting all the tokens and the associated POS labels repeatedly to ensure the suitability of the POS per word. The combined corpus is again used as the training corpus, and this process is repeated to ensure all data (words) are completely annotated. Figure 7 shows the operation of the Malay QTAG POS tagger.

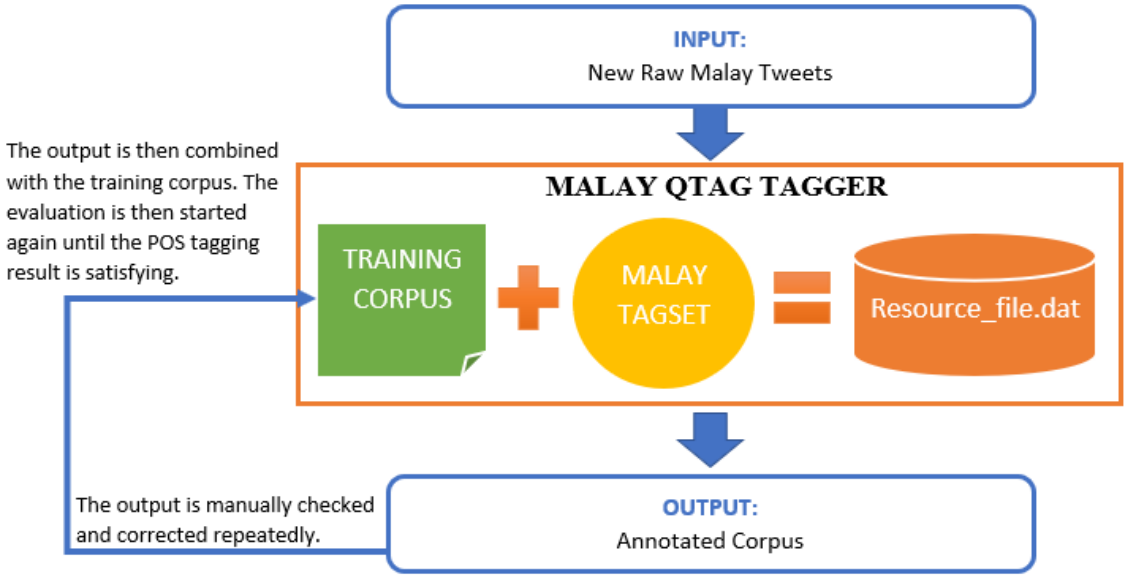

FIGURE 7. Illustration of the Malay QTAG POS tagger

RESULT

This section explains the evaluation phase of the Malay QTAG POS tagging. The evaluation phase is setup to evaluate the suitability of the Malay POS tag set and the developed training corpus with the raw Malay tweets taken from twitter.com. This phase requires two types of datasets (corpus): (i) the unnormalised (raw) test dataset, which is extracted directly from Twitter without changing the texts; and (ii) the normalised test dataset, which has the same corpus as that in (i) but has to undergo the normalisation process explained in the previous section (Figure 6).

Figures 8 and 9 show the evaluation of the new Malay tweets as the unnormalised (raw) test dataset by using the Malay QTAG POS tagger.

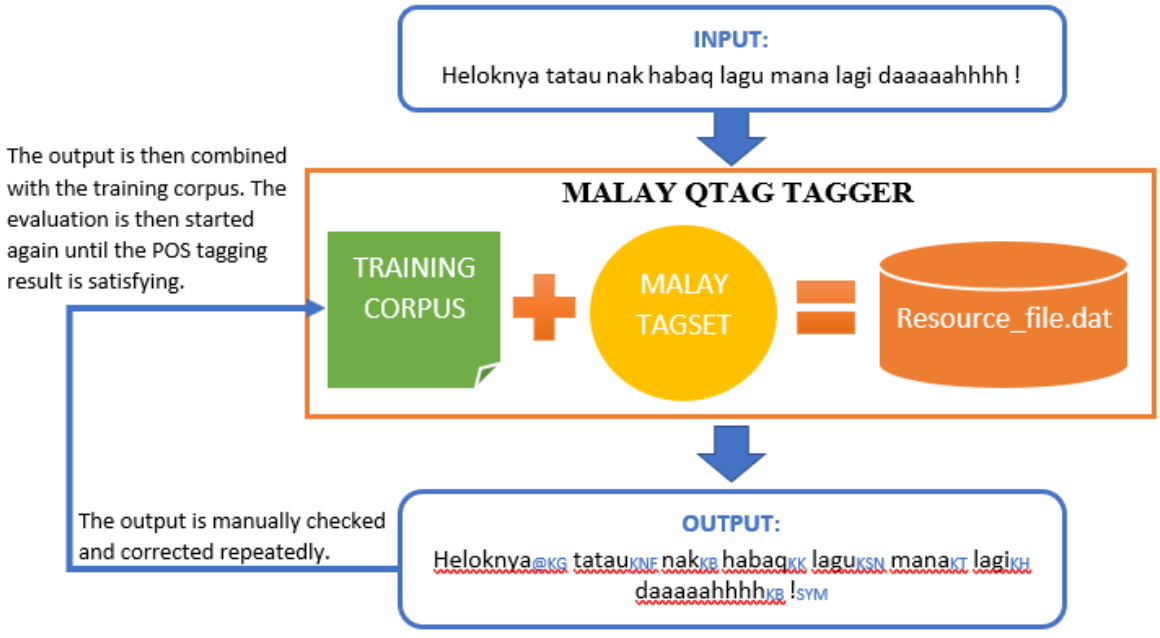

FIGURE 8. First example of evaluation phase of Malay QTAG POS tagger for the unnormalised dataset

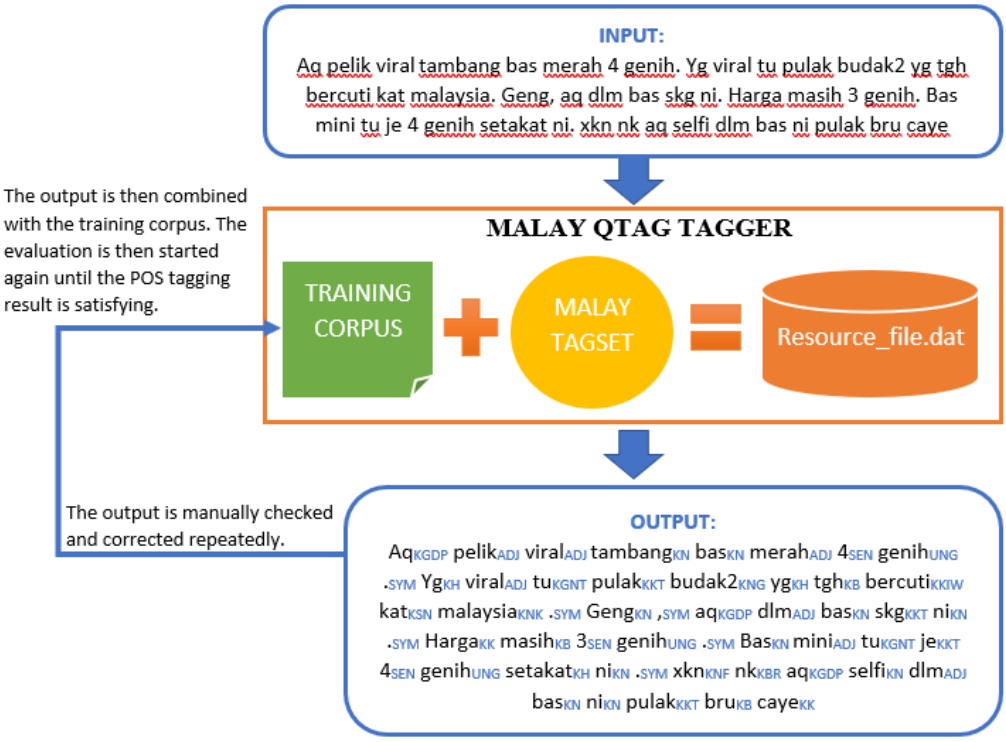

FIGURE 9. Second example of evaluation phase of Malay QTAG POS tagger for the unnormalised dataset.

Next, Figures 10 and 11 show the evaluation of the new Malay tweets as the normalised (undergone normalisation) test dataset by using the Malay QTAG POS tagger.

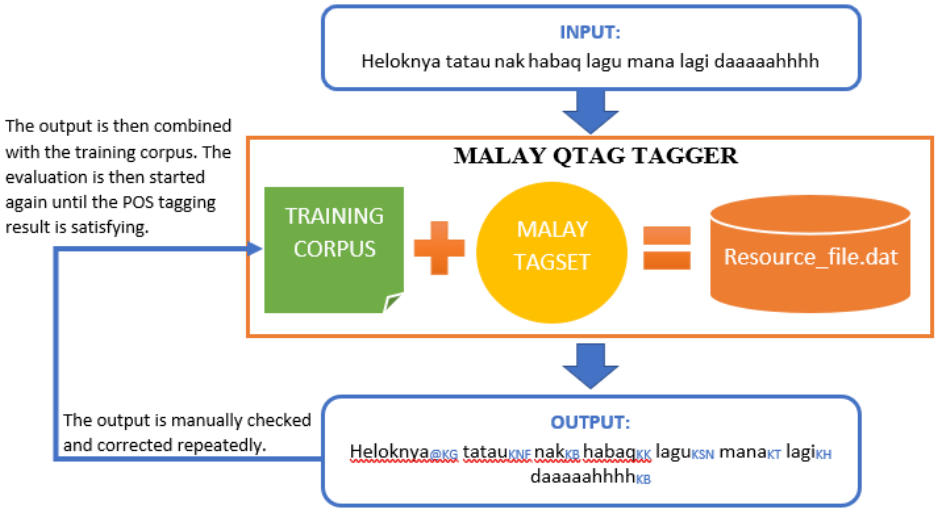

FIGURE 10. First example of evaluation phase of Malay QTAG POS tagger for the normalised dataset

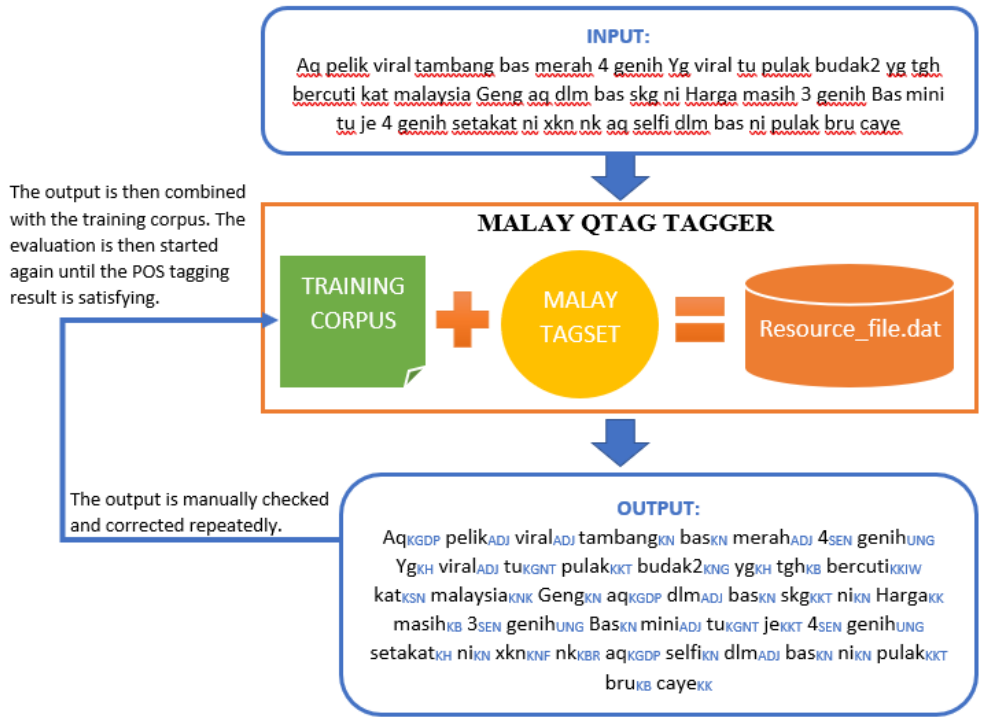

FIGURE 11. Second example of evaluation phase of Malay QTAG POS tagger for the normalised dataset

As mentioned earlier, the approach to evaluate the Malay POS tagging performance is based on the method used by Tufis and Mason (1998). The proposed evaluation is carried out by splitting the test data into small sets of test data and the accuracy obtained from all the small datasets. The performance is then assessed by calculating the average of all the obtained accuracies from the small test datasets.

Evaluation is carried out 15 times, and both test datasets (unnormalised and normalised test datasets) are divided into 15 small datasets. The results of the first set of test dataset (unnormalised test dataset) are shown in Table 4. The results of the evaluation of the second set of test corpus (normalised test dataset) are shown in Table 5. The overall evaluation of both test datasets is shown in Table 6.

TABLE 4. Results of unnormalised test datasets

Unnormalised Datasets	No. of Tokens	No. of Errors	Accuracy
CWS_1	**11**	**3**	**72.7%**
CWS_2	**141**	**5**	**96.4%**
CWS_3	305	18	94.1%
CWS_4	134	18	86.6%
CWS_5	287	33	88.5%
CWS_6	29	6	79.3%
CWS_7	130	5	96.2%
CWS_8	123	5	96.0%
CWS_9	155	5	88.4%
CWS_10	53	11	79.2%
CWS_11	140	21	85.0%
CWS_12	90	15	83.3%
CWS_13	**177**	**46**	**74.0%**
CWS_14	163	20	87.7%
CWS_15	174	25	85.6%

TABLE 5. Results of normalised test dataset

Normalised Datasets	No. of Tokens	No. of Errors	Accuracy
COS_1	**11**	**0**	**100.0%**
COS_2	**124**	**5**	**96.0%**
COS_3	301	5	98.7%
COS_4	133	4	97.0%
COS_5	289	11	96.2%
COS_6	29	1	96.6%
COS_7	127	3	97.6%
COS_8	124	3	97.6%
COS_9	157	4	97.5%
COS_10	56	2	96.4%
COS_11	138	15	89.1%
COS_12	88	13	85.2%
COS_13	**176**	**36**	**80.1%**
COS_14	164	2	98.8%
COS_15	172	18	89.5%

As shown in Table 4, the highest accuracy of 96.4% is obtained for the second test file (CWS_2) that contains 141 tokens with only five numbers of incorrectly labelled POS tags. Nearly the same accuracy (96.0%, Table 5) is obtained for the same test file that has been normalised (COS_2) before being tagged with POS. However, this situation does not apply to all the small dataset files. The test file CWS_1 in the unnormalised dataset (Table 4) achieves the lowest accuracy of only 72.7%. On the contrary, the same file that underwent normalisation (COS_1) achieves the highest accuracy. The best accuracy obtained is 100%, which indicates no error. Therefore, the average accuracy of the overall datasets can be used to evaluate the entire aspect of the test dataset (number of words and various styles of writings). The overall evaluation of accuracy on all the test files is conducted and presented in Table 6.

TABLE 6. Overall evaluation result on both test datasets

Test Datasets	No. of Tokens	No. of Errors	Accuracy
Unnormalised Datasets	2112	236	88.8%
Normalised Datasets	2089	122	94.6%

The overall evaluation results on these test datasets show that high average accuracy rates of more than 85% (88.8% and 94.6% for unnormalised and normalised datasets) are obtained. As shown in Table 6, the number of tokens in unnormalised datasets is higher than that in the normalised datasets because some tokens are symbols and punctuations. As described earlier, these tokens are removed during the normalisation process. The reduction in the number of errors from 236 to 122 (nearly 50% of error reduction) implies that the removal of symbols and punctuation plays an important role in improving the accuracy of the POS tagging on Malay social media texts.

DISCUSSION

A few points from the findings are worthy of discussion: (i) the impact of the simple normalisation process before the POS tagging, (ii) the capability of probabilistic approach (i.e. QTAG) (Tufis & Mason 1998; Mason & Tufis, 1997) when used for the Malay language (specifically on Malay social media texts) and (iii) the application of QTAG on the Malay language compared with the current state-of-art Malay POS tagging.

The higher accuracy on normalised test dataset than on the unnormalised test dataset may be due to the filtering of punctuation and symbols from the text. We presume that the punctuations and symbols contribute to the change of meaning, of a sentence or a word in the context of Malay social text writing. The change of meaning results in incorrect POS labelling.

The best evaluation result of the Malay QTAG POS tagging is an average accuracy of 94.6% (Table 6), which is considered a good result. The result is comparable to other adopted QTAG languages. For example, the Romanian QTAG achieves the best result of 98% for formal Romanian texts, and the German QTAG (Cox, 2010) obtains the best result of 90% and above for normalised German texts. Most previous works on QTAG (Cox, 2010; Tufis & Mason, 1998; Mason & Tufis, 1997) are based on formal texts. Conducting POS tagging of words with incorrect lexicon, structural errors, spelling mistakes and acronyms in social media, such as tweets, is challenging.

The performance of the Malay QTAG, which is built for automatic POS tagging of social media texts, is tested by comparing it with QTAG for other languages. Most previous works use supervised ML and semi-manually build the annotated corpus. Although the corpus is built in a similar manner in the current study, the range of informal text of POS tagging performance is from 74.28% (Albogamy & Ramsay, 2015) to 97.6% (Al-Sabbagh & Girju, 2012). Therefore, the Malay QTAG performs quite well given that its best result of accuracy is 94.6%.

Another contribution of the present study is the superiority of the Malay QTAG tagger to other automatic Malay POS taggers. Notably, the datasets used differ. Specifically, current works on Malay POS tagging are based on formal Malay texts, whereas this study focuses on informal social media texts. Thus, constructing a tagger exclusively for (informal) Malay POS tagging is more challenging than for formal Malay language texts. Despite the challenges, this study manages to obtain a high tagging accuracy of 94.6%, which is higher than that obtained by the novel work on formal Malay POS tagging by Halid and Nazlia (2017) with only 93.06%. The accuracy of formal text Malay POS tagging with various kinds of approaches is from 80% (Rayner et al., 2013) to 99.23% (Mohamed et al., 2015). In the present study, the best result is 94.6%. Therefore, the Malay QTAG performs well because its result is within the range of other automatic Malay POS taggers. The size of the annotated POS corpus influences the performance of supervised POS tagging, and this case is particularly true for the QTAG POS tagging approach (Cox, 2010). The total number of words in this study is only 5000, but QTAG can still perform with the best accuracy of 94.6% (Table 3). Mason and Tufis (1997) used a much larger scale of corpus with 23,000 words extracted from books. Therefore, the reported accuracy is higher with 98%. Accordingly, a definite future direction of this study is to work with a considerably large corpus for the optimum performance of the Malay QTAG.

The best performance of Malay POS tagging is 99.23% (Mohamed et al., 2015) using the SVM technique. This result implies that the performance of Malay POS tagging can be improved by applying a suitable supervised ML approach. Therefore, another future direction is to use automatic Malay POS tagger for social media texts by using other popular supervised approaches, particularly SVM.

CONCLUSION

A Malay QTAG POS tagger for social media texts has been successfully developed with a training corpus set containing over 5000 words and obtains an average Malay POS tagging accuracy of over 90%. The performance of this (informal) Malay QTAG POS tagger can be improved by increasing the size of the training corpus from 5000 words to 20,000 or 30,000

words. The tagger can also be improved by designing a detailed Malay POS tag set that is appropriate for informal Malay tweets. For example, all the slang words, such as *kud*, *la* and *ea*, in this study are classified by the UNG (idiom phrase POS tag) POS tag rather than its own POS tag set. This kind of changes can be made by editing the developed training corpus manually through searching the correspondent word and training the corpus again with the Malay QTAG tagger by using the command stated in the previous section.

Although this study has over 5000 words in the training corpus and achieves a POS tagging accuracy of above 90%, another improvement should be made by adding dialect words, such as Johor, Kelantan, Terengganu and other Malay dialect words, into the training corpus. A large number of dialect words put as input in the training indicate a high POS tagging accuracy on the informal Malay language. Thereafter, various types of supervised ML techniques can be applied. This approach will enable researchers to identify the best technique for the analysis of Malay POS tagging. Various kinds of supervised ML techniques can be applied to fully utilise the advantage of the Malay POS tagging. Given that POS tagging is an NLP application tool, future study will investigate the impact of this Malay POS tagging in sentiment analysis of social media texts in the informal Malay language.

ACKNOWLEDGMENT

We thank Dr Oliver Mason for providing us the program code as well as the consultation regards to program. This project is funded by MoHE under research code: FRGS/1/2016/ICT02/UKM/02/14.

REFERENCES

Abdulkareem, M. & Sabrina Tiun. (2017). Comparative Analysis of ML POS on Arabic Tweets. *Journal of Theoretical & Applied Information Technology. Vol. 95*(2), 403-411.

Albogamy, F. & Ramsay, A. (2015). POS Tagging for Arabic Tweets. International Conference Recent Advances in Natural Language Processing Proceedings, 7–9 September, Hissar, Bulgaria.

Al-Sabbagh, R. & Girju, R. (2012). A Supervised POS Tagger for Written Arabic Social Networking Corpora. KONVENS 2012 Conference Proceedings*,*19-21 September, Vienna, Austria.

Alshaikhdeeb, B. & Ahmad, K. (2016). Biomedical Named Entity Recognition: A Review. *International Journal on Advanced Science, Engineering and Information Technology*. *Vol.6* (6), 889-895.

Altawaier, M. M. & Sabrina Tiun, (2016). Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis. *International Journal on Advanced Science, Engineering and Information Technology*. *Vol. 6*(6), 1067-1073.

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.* Vol. 14(9), *843-851.*

Antony, P. J., Mohan, S. P. & Soman, K. P. (2010, March). SVM Based Part of Speech Tagger for Malayalam. 2010 International Conference in Recent Trends in Information, Telecommunication and Computing Proceedings, 12-13 March, Kerala, India.

Arbak Othman. (2005). *Kamus Komprehensif Bahasa Melayu*. Shah Alam: Oxford Fajar.

Berger, A., L., Della Pietra, SA. & Della Pietra, V.J. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics. Vol.* 22(1), 39-72.

Chekima, K. & Rayner Alfred. (2017). Sentiment Analysis of Malay Social Media Text. 4th International Conference on Computational Science and Technology Proceedings, 29-30 November, Kuala Lumpur, Malaysia.

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science And Technology. Vol. 37*(1), 51-89.

Cox, C. (2010). Probabilistic Tagging of Minority Language Data: A Case Study Using Qtag In Gries, T. S., Wulff, S. & Davies, M. (Eds.), *Corpus Linguistic Applications: Current Studies, New Directions* (pp. 213-231). Amsterdam: Rodopi.

Derczynski, L., Ritter, A., Clark, S. & Bontcheva, K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. International Conference Recent Advances in Natural Language Processing RANLP 2013 Proceedings, 7-13 September, Hissar, Bulgaria.

Elworthy, D. (1995). Tagset Design and Inflected Languages. Paper presented at EACL SIGDAT Workshop. Dublin, Ireland, January.

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. 49th Annual Meeting of the Association for Computational Linguistics Proceedings, 19-24 June, Portland, Oregon.

Gui, T., Zhang, Q., Huang, H., Peng, M. & Huang, X. (2017). Part-of-speech Tagging for Twitter with Adversarial Neural Networks. 2017 Conference on Empirical Methods in Natural Language Processing Proceedings, 7-11 September, Copenhagen, Denmark.

Halid, N. A. & Nazlia Omar. (2017). Malay Part of Speech Tagging Using Ruled-Based Approach. *Asia-Pacific Journal of Information Technology and Multimedia*. *Vol. 6*(2), 91-107.

Hassan Ahmad. (1985). The Role of Dewan Bahasa dan Pustaka in the Advancement of Indigenous Academic Publishing in Malaysia. In S. Gopinathan (Ed.), *Academic Publishing in ASEAN*. Singapore: Festival of Books Singapore.

Hawkins, J. M. (2008). *Kamus Dwibahasa Bahasa Inggeris–Bahasa Malaysia*. Selangor: Oxford Fajar.

Hock, O. Y. (2009). *Kamus Dwibahasa*. Petaling Jaya: Pearson Longman.

Java, A., Song, X., Finin, T. & Tseng, B. (2007, August). Why We Twitter: Understanding Microblogging Usage and Communities. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis Proceedings, 12-15 August, San Jose, CA, USA.

Juhaida Abu Bakar, Khairuddin Omar, Mohammad Faidzul Nasrudin & Mohd Zamri Murah. (2013). Morphology Analysis in Malay POS Prediction. International Conference on Artificial Intelligence in Computer Science and ICT (AICS 2013) Proceedings, 25-25 November, Langkawi, Malaysia.

Juhaida Abu Bakar, Khairuddin Omar, Mohammad Faidzul Nasrudin & Mohd Zamri Murah. (2013). Part-of-Speech for Old Malay Manuscript Corpus: A Review. Second International Multi-Conference on Artificial Intelligence Technology (M-CAIT'13) Proceedings, 28-28 August, Shah Alam, Malaysia.

Karim Harun & Maslida Yusof. (2015). Komunikasi Bahasa Melayu-Jawa Dalam Media Sosial. *Jurnal Komunikasi, Malaysian Journal of Communication*. *Vol.31*(2), 617-629.

Knowles, G. O. & Zuraidah Mohd. Don. (2006). *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Li, D. (1998) The Plight of the Purist. In Pennington, M. (Ed.). *Language in Hong Kong at Century's End* (pp. 161-190). Hong Kong: Hong Kong University Press.

Maslida Yusof (2018). Trend Ganti Nama Diri Bahasa Melayu dalam Konteks Media Sosial. *Jurnal Komunikasi, Malaysian Journal of Communication*. *Vol. 34*(2), 36-50.

Mason, O. & Tufis, D. (1997). Probabilistic Tagging in a Multi-Lingual Environment: Making an English Tagger Understand Romanian. Third European TELRI Seminar proceedings, 16-18 October, Montecatini, Italy.

Mohamed, H., Nazlia Omar & Mohd. Juzaiddin Ab. Aziz. (2015). Malay Part of Speech Tagger: A Comparative Study on Tagging Tools. *Asia-Pacific Journal of Information Technology and Multimedia*. *Vol. 4*(1), 11-23.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. United Kingdom: Cambridge University Press.

Nasiroh Omar, Ahmad Farhan Hamsani, Nur Atiqah Sia Abdullah & Siti Zaleha Zainal Abidin. (2017). Construction of Malay Abbreviation Corpus Based on Social Media Data. *Journal of Engineering and Applied Sciences*. *Vol. 12*(3), 468-474.

Nguyen, T. M. H., Vu, X. L. & Le-Hong, P. (2003). A Case Study of the Probabilistic Tagger QTAG for Tagging Vietnamese Texts. 1st National Conference ICT RDA Conference Proceedings.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *The Computing Research Repository (CoRR11)*. http://dblp.uni-trier.de/db/journals/corr/corr1103.html#abs-1103-2903.

Nøklestad, A. & Søfteland, Å. (2007). Tagging a Norwegian Speech Corpus. 16th Nordic Conference of Computational Linguistics NODALIDA-2007 Proceedings, 25-26 May, Estonia, Tartu.

Nooralahzadeh, F., Brun, C. & Roux, C. (2014). Part of Speech Tagging for French Social Media Data. 25th International Conference on Computational Linguistics Conference (COLING 2014) Proceedings, 23-29August, Dublin, Ireland.

Nurul Iman Ahmad Bukhari, Azu Farhana Anuar, Khairunnisa Mohad Khazin & Tengku Mohd Farid Bin Tengku Abdul Aziz. (2015). English-Malay Code-Mixing Innovation In Facebook Among Malaysian University Students. *International Refereed Research Journal*. *Vol. 6*(4), 1-10.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N. & Smith, N. A. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings, 9–14 June, Westin Peachtree Plaza Hotel Atlanta, Georgia, USA.

Rayner Alfred, Adam Mujat & Joe Hendry Obit. (2013). A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles. Asian Conference on Intelligent Information and Database Systems Proceedings, 18-20 March, Kuala Lumpur, Malaysia.

Sornlertlamvanich, V., Charoenporn, T., & Isahara, H. (1997). ORCHID: Thai Part-Of-Speech Tagged Corpus. Technical Report, National Electronics and Computer Technology Center .

Tagging with QTAG. (2007). Retrieved 15 May, 2018 from https://www1.essex.ac.uk/linguistics/research/resgroups/clgroup/Resources/Nugues/QTAG/

Toutanova, K. & Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. 2000 Joint SIGDAT Conference on

Empirical Methods in Natural Language Processing and Very Large Corpora Proceeding, 7-8 October, Hong Kong.

Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technolog Proceeding, 27 May - 1 June, Edmonton, Canada.

Tran, O. T., Le, C. A., Ha, T. Q. & Le, Q. H. (2009). An Experimental Study on Vietnamese POS Tagging. International Conference Asian Language Processing (IALP'09) Proceedings, 7-9 September, Singapore.

Tufis, D. & Mason, O. (1998, May). Tagging Romanian Texts: A Case Study for Qtag, A Language Independent Probabilistic Tagger. First International Conference on Language Resources and Evaluation (LREC) Proceedings, 28-30 May, Granada, Spain.

van der Goot, R., Plank, B. & Nissim, M. (2017). To Normalize, or Not to Normalize: The Impact of Normalization on Part-of-Speech Tagging. 3rd Workshop on Noisy User-generated Text Proceedings, 7 September, Copenhagen, Denmark.

Xian, B. C. M., Lubani, M., Ping, L. K., Bouzekri, K., Mahmud, R. & Lukose, D. (2016). Benchmarking Mi-Pos: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering*. *Vol*. *2*(3), 115-121.

Yang, L. C., Selvaretnam, B., Hoong, P. K., Tan, I. K., Howg, E. K. & Kar, L. H. (2016). Exploration of Road Traffic Tweets for Congestion Monitoring. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*. *Vol. 8*(2), 141-145.

ABOUT THE AUTHORS

Siti Noor Allia Binti Noor Ariffin: An undergraduate degree student at the Faculty of Information Science and Technology (FTSM) in Universiti Kebangsaan Malaysia (UKM). Currently, she is under the supervision of Dr. Sabrina Tiun for an industrial training in research track program.

Dr. Sabrina Tiun: A senior lecturer at the Faculty of Information Science and Technology (FTSM) in Universiti Kebangsaan Malaysia (UKM). Range of research interests are form Natural Language Processing to Speech Processing and Information Retrieval. Member of the Knowledge Computing research group, under the Centre of Artificial Intelligence.

eISSN: 2550-2131
ISSN: 1675-8021