Improved POS Tagging Model for Malay Twitter Data based on Machine Learning Algorithm

Siti Noor Allia Noor Ariffin[1], Sabrina Tiun[2]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*—Twitter is a popular social media platform in Malaysia that allows for 280-character microblogging. Almost everything that happens in a single day is tweeted by users. Because of the popularity of Twitter, most Malaysians use it daily, providing researchers and developers with a wealth of data on Malaysian users. This paper explains why and how this study chose to create a new Malay Twitter corpus, Malay Part-of-Speech (POS) tags, and a Malay POS tagger model. The goal of this paper is to improve existing Malay POS tags so that they are more compatible with the newly created Malay Twitter corpus, as well as to build a POS tagging model specifically tailored for Malay Twitter data using various machine learning algorithms. For instance, Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) classifiers. This study's data was gathered by using Twitter's Advanced Search function and relevant and related keywords associated with informal Malay. The data was fed into machine learning algorithms after several stages of processing to serve as the training and testing corpus. The evaluation and analysis of the developed Malay POS tagger model show that the SVM classifier, as well as the newly proposed Malay POS tags, is the best machine learning algorithm for Malay Twitter data. Furthermore, the prediction accuracy and POS tagging results show that this research outperformed a comparable previous study, indicating that the Malay POS tagger model and its POS were successfully improved.**

Keywords*—*Informal Malay; Malay Twitter corpus; Malay POS tagging; Malay POS tagger model, Malay social media texts; Malay POS machine learning

I. Introduction

In general, POS is a classification system for words that classifies them according to their usage and function in sentences [1]. Similarly, a POS tagger is a component of software that reads the text in multiple languages and assigns appropriate words to each word (or other tokens) in the text. Malay has four leading POS tags, namely nouns, verbs, adjectives, and word tasks [2]; however, these leading POS tags are more suitable for tagging standard Malay text than informal Malay text such as Malay Twitter data.

Twitter is a microblogging service that combines social networking websites and instant messaging technologies to create a network of users from all walks of life who can communicate throughout the day via 280-character [3][4] short messages called tweets. Additionally, users can use Twitter's Trending features to follow specific topics. Tweets can range from jokes to current events to dinner plans, yet they cannot exceed Twitter's characters limit [5].

Twitter's Application Programming Interface (API) simplifies the process of collecting Twitter data. However, Twitter imposed specific fees and requirements for data access. Clearly, according to [6], collecting tweets using a standard Twitter API account is limited to the most recent seven days of data, and collecting and accessing tweets older than those seven days requires a premium Twitter API account efficiently cost hundreds of dollars. On top of that, Twitter provides an Advanced Search feature that enables users to refine search results based on date ranges, people, and more. Hence, this study collects tweets written in informal Malay, incorporates various dialects, conversational slang languages, and mixed languages via Twitter's Advanced Search feature [1][6][7], rather than the costly and limited API [6]. However, this data collection limited to the words contained in tweets due to other tweet features, such as user information (full name & username), hashtags, URLs, and timestamps, are considered superfluous. Hence, this study purposefully ignored and omitted it.

Malay (ISO 639-3; MSA) is a formal language spoken by Malaysians, Indonesians, Singaporeans, and Bruneians of all races. On the contrary, informal Malay is a dialect of Malay that Malaysians use in everyday conversation. Informal Malay encompasses a diverse range of informal terms, such as accent (or dialect) words, slang, titles (e.g., hang, mek), sounds (such as words written to express sounds like laughter, cat sounds, and knocking sounds), and mixed languages. Firstly, the term 'regional dialect or language' refers to a group of people who speak a country state's language, resulting in word variation. In comparison, slang is understood by a minuscule percentage of the population. Following that, 'mixed language' refers to the simultaneous use of a foreign language and Malay. For instance, when users write on social media to share feedback, express an opinion or stories, they frequently use every day conversational language to convey a friendly, casual, and easy-going image to other users.

Moreover, numerous research papers have been published in recent years on the Malay Twitter corpus and the prediction and POS tagging of Malay Twitter data; however, there has been a dearth of information in various areas, which requires improvement. For instance, Malay Twitter data normalization techniques [5], Malay POS tagger explicitly tailored for such data [1], and supervised machine learning algorithms that best suit the mentioned data. According to [8], one way for improving the quality of language processing on social media data is to automate non-standard terms to their corresponding standard tokens using normalization methods. Hence, this study chose to utilize [5]'s improvised normalization

techniques for Malay Twitter data, dubbed Malay Text Normalizer. Furthermore, their Malay Text Normalizer demonstrated acceptable performance with POS tagging on a normalized tweets test corpus.

Besides, Malay Twitter data can be tagged with POS in numerous ways. Using supervised machine learning algorithms to build a POS tagging model is one of the most widely used methods. According to [1], developing a Malay POS tagging model specifically for Malay Twitter data is difficult due to the presence of dialects, grammatical and typographical errors, and abbreviations. Nevertheless, they have successfully developed a Malay POS tagger that can tag Malay Twitter data using a supervised machine learning algorithm, QTAG, a language-independent probabilistic tagger by [9]. In addition, their Malay QTAG tagger produces exceptional results on both normalized and unnormalized test corpora. Therefore, in this study several different supervised machine learning algorithms such as SVM, NB, DT, and KNN classifiers will be employ. This technique has been applied to a wide variety of research fields. The primary reason for using these supervised machine learning algorithms in this study is that it has demonstrated satisfactory results in other languages and is underutilized in informal Malay Twitter data. Thus, it is critical to investigate how well these supervised machine learning algorithms perform on such data.

This study aims to enhance existing Malay POS tags by [1] to make them more compatible with the newly created Malay Twitter corpus and develop a POS tagging model specifically for the Malay Twitter corpus using the previously mentioned machine learning algorithms. Throughout this paper, this study has successfully made several significant contributions, such as collecting and extracting Malay Twitter data by employing informal Malay terms as keywords, tagging the Malay Twitter corpus with newly proposed improvised Malay POS tags, and analyzing the corpus data with the newly developed data Malay POS tagging model. Nonetheless, until the data's copyright is enforced, this Malay Twitter data collection will be inaccessible to the public or future research.

The following is how the rest of this paper is organized: Section II summarizes relevant works, Section III discusses the methodology of this study, Section IV discusses the discussion of this study, Section V presents the results and analysis, and Section VI summarizes the study as well as several suggested future works.

II. Related Work

Malay POS tags are a type of Malay POS used to tag words in this language. As stated previously, Malay has four primary POS tags; however, this POS tag is insufficient for tagging words in the Malay Twitter corpus. For this reason, this study decided to create new Malay POS tags by referencing Malay formulas and grammar by [2]. Furthermore, this study modified the newly created POS tags to meet the Malay Twitter data criteria by comparing the word classes discovered in [2] and [10], as well as some previous findings on social media texts [1].

Additionally, a study by [11] created several new POS tags to meet their research requirements. This study discovered that several of [11]'s newly developed POS tags are suitable for categorizing words in the Malay Twitter corpus. Consider the FOR and NEG POS tags, for example. The FOR POS tag is used to classify words in foreign languages found in the study corpus. Similarly, the NEG POS tag identifies words with negative connotations, such as those that refer to swearing. Therefore, this study chose to use [11]'s two newly developed POS tags, namely FOR and NEG POS tags, as the Malay Twitter data writing demonstrates that Malaysians enjoy combining words from multiple languages in tweets and using negative words to express emotions or disapproval of situations. In addition, removing a foreign language from the corpus alters the author's intended meaning and the structure of the tweets' writing. This change will result in an error when auto annotations and annotators attempt to tag the POS tags. In that case, this study will not exclude foreign language terms, slang terms, or informal Malay expressions that adhere to this principle—instead, a unique POS tag designed for this type of word tagging. Moreover, by referencing to [11], this study generated several new Malay POS tags, namely LD, SL, GL, and BY POS tags. Firstly, the LD POS tag is used to indicate words with an accent (or dialect). Secondly, the SL POS tag is used to indicate slang language. Following that, the GL POS tag is used to indicate words referred to nicknames, and lastly, the BY POS tag used to indicate words that express sounds. Besides, this study also combined two POS tags that contained the same word into a single POS tag, such as the GDT-KTY POS tag, which comprises the POS tag for self-query pronouns (kata ganti nama diri tanya) and the POS tag for query word (kata tanya). Finally, another newly developed Malay POS tag was created by combining several prominent POS tags with sub-POS tags for shortened words, accents (or dialects), and negative particles. For instance, the term 'iols' is an abbreviation for a foreign language term. Therefore, the appropriate POS tag for this term combines the foreign language primary POS tag (FOR) and the shortened word sub-POS tags (KEP).

Malaysians, particularly teenagers, are incredibly inventive in writing and developing new words [12]. As a result, Malay social media text, such as Malay tweets, is saturated with informal Malay and peppered with mixed-language phrases and derogatory terms. POS tagging is exceptionally complicated, time-consuming, and energy-intensive for this type of corpus. If a word is not suitable to any existing POS tags, researchers will need to find alternative initiatives to either remove the word or tag it with any existing POS tags that they deem appropriate. This technique, however, is impractical due to the researchers' manipulation of study data. Thus, this study took the initiative to create new Malay POS tags that are compatible with Malay Twitter data to avoid confusion and expedite the POS tagging process; this study not only added and used the FOR and NEG POS tags from [11] and created several new Malay POS tags, but also added six additional Malay POS tags from [1]. With these numerous newly crafted Malay POS tags, this study's total number of POS tags has increased to 45 tags. As stated earlier, Malay POS tags were explicitly created to tag words in Malay Twitter data, saving researchers and annotators time to tag words with the correct POS tags.

This study conducted prediction and Malay POS tagging using the Malay POS tagger model developed based on four different machine learning algorithms from similar past studies [1], namely SVM, NB, DT, and KNN classifiers. Firstly, SVM is a widely used classifier in this type of research using machine learning algorithms, as it can accurately predict and tag POS tags [13]. Additionally, SVM predicts the POS tags for unknown words [14][15] and is considered one of the most efficient and accurate classification techniques for POS tagging [16]. According to [14], this classifier is based on sub-words, contextual information, and environmental context tagger, and [16] noted that this classifier could be used to classify sentences, ambiguity classes, and word length. Secondly, the NB classifier generates documents using the Bayes rule theory. NB classifier is extensively used in research related to predicting and tagging POS tags [17], as it can estimate the probability of each word feature before building the classification model. Five classification techniques are included in the NB class: Gaussian, Multinomial, Complement, Bernoulli, and Categorical. Following that, DT is a highly simple-to-understand and interprets machine learning class. DT works by developing predictive models for target variables. This classifier requires only a tiny amount of configuration data to operate and supports two distinct data types: numeric and categorical. DT can be validated using statistical tests and perform well, even if the data model occasionally rejects the resulting prediction. This classifier can overcome the information breakdown problem by obtaining accurate estimates of the probability of change [18] and can be used to tag unknown words with POS tags based on their word endings, word forms, and context information. Lastly, because the KNN classifier relies on POS tags associated with the test corpus, it does not generate a clear declaration representation; however, it estimated the new word POS tag by comparing it to words in the training set.

Generally, this study evaluated and quantified the prediction accuracy and POS tagging using these four-evaluation metrics: precision, recall, F1-score, and accuracy. Firstly, the precision evaluation metrics quantify machine learning algorithms' ability to avoid predicting and tagging negative samples as positive. Secondly, the recall evaluation metric assesses a machine learning algorithm's ability to re-identify positive samples. These terms refer to whether the predictions made by such machine learning algorithms are appropriate for external assessment or not. Following that the F1-score evaluation metrics can be interpreted as a weighted harmonic mean for precision and recall evaluation metrics. Furthermore, this study discovered that Malay social media text's prediction accuracy and POS tagging could be evaluated and quantified using all four-evaluation metrics. In addition, this study also discovered that the computation value for each evaluation metric could be easily generated via the classification report's evaluation metrics [19]. Classification reports are evaluation metrics that present computation values for each of the four-evaluation metrics in a classification report table. Thus, this study chose to evaluate and quantify the prediction accuracy and POS tagging using the sklearn library's evaluation metrics for classification reports [20].

III. Methodology

This study aims to improve existing Malay POS tags to fit the new Malay Twitter corpus better and develop a tagging model tailored to the new Malay Twitter corpus using the previously mentioned machine learning algorithms. The proposed methodology entails data collection and pre-processing as shown in Fig. 1. As stated previously, this study's data was compiled using Twitter's Advanced Search feature. It conducts the search using the provided keywords. Pre-processing of data includes data normalization and annotation. After that, vectorization is then applied to the data, preparing it for use by the machine learning algorithms: SVM, NB, DT, and KNN classifiers. This study vectorizes the data using the sklearn library's vectorization techniques [20], converting them to numbers, as machine learning algorithms operate exclusively on numbers. Then, to effectively use these algorithms, they must be trained to extract both word characteristics and their POS from the data. This procedure generates a model upon which Malay tweets can be automatically tagged with the Malay POS tags based solely on their context. The following sections go into details about each process.

A. Data Collection

This study gathered data manually by focusing on keywords associated with informal Malay and restricting the date ranges to February 2019. The keywords were selected following a review of the literature on informal Malay and structure. Table I contains a sample of keywords derived from previous studies and used to collect data. As stated earlier, the standard method of collecting Twitter data is through their API, enabling developers and researchers to collect data quickly; however, the API comes with a slew of restrictions, including a seven-day limit on tweets and a limit on the number of requests made to the Twitter server [6]. As a result, this study chose to manually collect data using Twitter's Advanced Search feature, rendering the limitations mentioned previously obsolete.

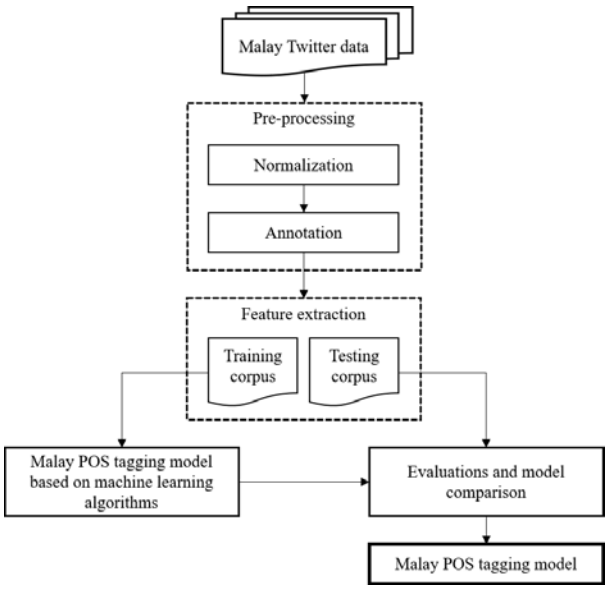

Fig. 1. Architecture of the Malay POS Tagging Model.

TABLE I. THE FOLLOWING ARE SOME OF THE KEYWORDS ASSOCIATED WITH SLANG TERMS [17] USED TO COLLECT DATA

Slang Word	Meaning in Malay	Meaning in English	Slang Form
bro	*awak, kamu, abang*	you, brother	Pronouns
gua	*saya, aku*	me	Pronouns
korang	*awak semua, kamu semua*	all of you, you all	Pronouns
pastu	*selepas itu*	after that	Expression
kekadang	*kadang-kadang*	sometimes	Expression
tetibe	*tiba-tiba*	out of the blue, suddenly	Expression
heran	*hairan*	astonished	Expression
takpe	*tidak mengapa*	it's okay	Expression
takleh	*tidak boleh*	can't	Expression
citer	*cerita*	story, tale	Expression
sesama	*bersama-sama*	together	Expression
jeles	*cemburu*	jealous	Foreign language
terer	*pandai*	clever	Foreign language
kaler	*warna*	colour, tone	Foreign language
hensem	*kacak*	handsome	Foreign language

This study focuses entirely on a single criterion for tweet inclusion: tweets must be written in informal Malay. As noted previously, informal Malay is replete with colloquial terms such as dialect, slang, titles, sounds, and mixed languages. Eventually, this study used only keywords derived from previous research findings to ensure that the chosen tweets were appropriate and accurately reflected this study's objective. On the contrary, there are currently no exclusion criteria for tweets as this study compiled a list of all tweets returned in response to the keywords entered. However, as explained earlier, several additional tweet features were purposefully overlooked and omitted, as the study's goal is to collect only informal Malay text. Therefore, this study ignores all other characteristics to focus exclusively on textual characteristics and the frequency with which informal Malay terms were used in social media texts.

B. Data Pre-processing

As previously stated, the data for this study will be pre-processed using two well-known pre-processing steps: data normalization and annotation. The pre-processing of the data begins with data normalization. The data normalization process strips Malay Twitter data of all ambiguous signs, symbols, and spellings. This normalization process is required for accurate POS tagging; however, accurate POS tagging is only possible after informal terms in the data are converted to their standard form (spellings). Therefore, this study's data were normalized using Malay Text Normalizer, a rule-based normalizer designed to normalize only Malay, Romanized Arabic, and English words [5] in the corpus. Subsequently, in the next pre-processing steps: data annotation, the normalized data will be annotated with POS. The data annotation process needs to ensure that each word in the data is appropriately tagged with the correct POS tags. For this reason, this study's data were manually annotated [1] using the newly proposed Malay POS tags. The final dataset contains 1,791 tweets in various languages relevant to informal Malay and related to the previously mentioned keywords.

C. Feature Extraction and Model Training

Following data collection and pre-processing, the final dataset is divided into two corpora: training and testing. The training corpus contains 70% of the data from the final dataset and is used to train these machine learning algorithms on extracting valuable features from words and their contexts using [21] features extraction method. This study extracted ten significant features, including the preceding and following words, the prefix of each word (limited to the first three alphabets of the word), the suffix of each word (limited to the last three alphabets of the word), the word's length, and the presence of a digit in the word. These features were extracted to aid the POS tagging model in automatically assign the correct POS to each word in the corpus, based on its context alone.

Additionally, as previously explained, the training corpus is vectorized before being fed into machine learning algorithms. Vectorization converts data from a textual to a numerical format, as these algorithms only work with numbers. Hence, this study leverages [20]'s sklearn library by vectorizing the training corpus using its vectorization technique. After that, all machine learning classifiers are trained using this vectorized training corpus by incorporating it into the classifier's code in the sklearn library. The sklearn library by [20] used in this study because it includes all necessary algorithms [22], such as machine learning classifiers and vectorization techniques, and evaluation methods. Besides, the library is simple to use, as the algorithms are invoked via the provided code. Finally, the accuracy of these machine learning classifiers will be evaluated using the sklearn library's evaluation metrics for classification reports and tested using the testing corpus, which contains the remaining 30% of the final dataset's data. Table II contains a summary of the features used in this phase.

TABLE II. LIST OF FEATURES USED IN THIS STUDY

Categories	Feature Description
Word Prefixes & Suffixes	Prefix 1 (first letter) Prefix 2 (first two letters) Prefix 3 (first three letters) Suffix 1 (last letter) Suffix 2 (last two letters) Suffix 3 (last three letters)
Token (Word) Context	The previous word Next word Word length Does the word contain digits

The features listed in Table II were chosen because:

1) According to [21], the word before it and its POS tag serves as a guideline for identifying the appropriate POS tag for the word after it.

2) Author in [21] also stated that the presented algorithm tags the current word with the appropriate POS tags based on the word information and surrounding POS tags.

3) The letter at the beginning and end of each word refers to the letter's size and position in the word. This feature, according to [23], has the potential to influence the effectiveness of POS tagging.

4) The length of each word is a binary feature that determines its size [23], [24].

5) The author in [23] defines digit features as features based on the presence of digits (numbers) and symbols in words.

6) According to [23] and [24], the context of words in a sentence influences the tagging value of unigrams in the corpus.

IV. Discussion

This study gathered tweets written in informal Malay from numerous Malaysian users through relevant and related keywords associated with informal Malay language. The keywords searched was done by using the same method used by [1][6][7] which is through Advanced Search Twitter function located at the Twitter main homepage. The collected data then was normalized using a Malay normalizer by [5] and annotated with the newly created Malay POS tags in the pre-processing process. Additionally, this study extensively used the sklearn library code by [20] to implement the classifiers algorithm, vectorization techniques, and evaluation method. Following that, this study vectorized the data to prepare it for the machine learning algorithms as the classifiers can only read data in numerical format. Finally, this study used the sklearn library [20] to develop a Malay POS tagger model based on these machine learning algorithms, namely SVM, NB, DT, and KNN, and evaluate each classifier's results.

V. Results and Analysis

This study developed a Malay POS tagger models using a training corpus and machine learning classifiers code from the sklearn library in previous sections. Therefore, this section will evaluate the developed Malay POS tagger models using the test corpus to determine its functionality. The evaluation of POS tagging by Malay POS tagger models employs metrics for classification reports from the sklearn library to determine the accuracy of POS tagging across all four machine learning algorithms. Table III summarizes the algorithms' evaluation results using the prepared testing corpus and the details mentioned previously.

Based on the table above, this study discovered that the SVM classifier achieved a relatively high predictive accuracy and Malay POS tagging at 94%, and the DT classifier had the second-highest score at a rate of 93%. Besides, the SVM and DT classifiers also rated the highest POS tagging evaluation results on the same scoring metric, i.e., F1-score (0.92 & 0.89). In other words, this evaluation demonstrates that both classifiers generate nearly identical predictive outcomes and POS tagging. The reason for this is that while the SVM classifier is computationally simple to implement, it has a high computational cost [14][15], whereas the DT classifier is capable of handling sparse data problems and still produces the best results when tested on small size data [25][26].

TABLE III. The Following Summarized the Classification Reports for POS Tagging Generated by All Four Machine Learning Classifiers

Machine Learning Classifier	Classification Reports			
	Precision	***Recall***	***F1-score***	***Accuracy***
Support Vector Machine (SVM)	0.93	0.91	0.92	0.95
Naïve Bayes (NB)	0.74	0.57	0.60	0.85
Decision Tree (DT)	0.89	0.88	0.89	0.93
K-Nearest Neighbor (KNN)	0.85	0.87	0.85	0.90

This study examined the algorithms' functionality using extensive data collection, resulting in up to 95% accuracy. Based on this result, the developed Malay POS tagging model is ready to predict and tag Malay Twitter data using the newly created Malay POS tags. The evaluation and analysis results indicate that the SVM classifier is the optimal machine learning algorithm for Malay Twitter data and that the Malay POS tags used are also optimal for such data. Furthermore, this study's prediction accuracy and POS tagging score are higher than those of a similar previous study, indicating that this study successfully improved the Malay POS tagger model and its POS.

A. Model Comparisons

As stated earlier, this study was based on an earlier study by [1]. The author in [1] investigated the prediction and tagging of Malay POS by developing Malay POS tagger models based on the Hidden Markov Model (HMM) trigram machine learning algorithm, the QTAG model. As a result, this study took the initiative to conduct comparable studies using various machine learning algorithms, including SVM, NB, DT, and KNN classifiers. The purpose of this study was to determine which machine learning algorithms are best suited for processing Malay POS predictions and tagging, particularly Malay Twitter data.

While both studies used the same data type, the pre-processing of the data was significantly different. For instance, this study normalized the data using the Malay Text Normalizer by [5], whereas [1] used only a few simple data pre-processing steps such as the removal of punctuation marks, symbols, and numbers. Following that, the Malay POS tags used in the annotation process of the two studies are distinct, as each study developed its own Malay POS tags that correspond to the content of its study data. Finally, by comparing the prediction accuracy and POS tagging results from these two studies, the results obtained by this study were significantly higher, with a difference of 0.29% from [1]. Eventually, the results and analysis demonstrate that the SVM classifier is the optimal machine learning algorithm for predicting and tagging Malay POS in Malay Twitter data and that the newly proposed Malay POS tags are appropriate for use as POS in such data. The distinction between this study and [1] is summarized in Table IV.

TABLE IV. THE FOLLOWING TABLE COMPARES [1] TO THIS STUDY

Studies	[1]'s study	This study
Objectives	Malay POS tagger	
Corpus	Informal Malay tweets	
Technique	Machine learning	
Algorithms	HMM (QTAG)	SVM, NB, DT, & KNN
POS Size	38	45
No. of Tweets	300	1,791
No. of Words	5,513	38,714
Accuracy	94.60%	95.00%

VI. CONCLUSION AND FUTURE WORK

In general, this study aims to improve on [1], work by developing a POS tagging model tailored to Malay Twitter data using a variety of machine learning algorithms, including SVM, NB, DT, and KNN classifiers. The collected data, as well as the newly developed Malay POS tags and Malay POS tagger model, are expected to be of assistance to researchers and developers, particularly those with expertise in informal Malay and related Natural Language Processing fields. This study can be improved further by collecting more non-standard Malay words in the training corpus, preferably more than 10,000 tweets specific to a specific domain such as food or health, rather than general domain. This enhancement is proposed to ensure that the training corpus contains enough data for the subsequent processing process.

ACKNOWLEDGMENT

This research work is funded by The Ministry of Higher Education Malaysia under research grant code: FRGS/1/2020/ICT02/UKM/02/1.

REFERENCES

[1] Ariffin, S. N. A. N., & Tiun, S. "Part-of-Speech Tagger for Malay Social Media Texts". GEMA Online® Journal of Language Studies, 18(4), 2018.

[2] Safiah, K. N., Onn, F. M., Musa, H. H., & Mahmood, A. H. "Tatabahasa Dewan Edisi Ketiga". Kuala Lumpur: Dewan Bahasa dan Pustaka, 2010.

[3] Meftah, S., & Semmar, N. "A neural network model for part-of-speech tagging of social media texts". In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018.

[4] Kumar, P., & Gruzd, A. "Social Media for Informal Learning: a Case of# Twitterstorians". In Proceedings of the 52nd Hawaii International Conference on System Sciences, January 2019.

[5] Ariffin, S. N. A. N., & Tiun, S. "Rule-based text normalization for Malay social media texts". International Journal of Advanced Computer Science and Applications, 11(10), 2020.

[6] Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., & Hazim, M. "Halal products on Twitter: Data extraction and sentiment analysis using a stack of deep learning algorithms". IEEE Access, 7, 83354-83362, 2019.

[7] Izazi, Z. Z., & Tengku-Sepora, T. M. "Slangs on Social Media: Variations among Malay Language Users on Twitter". Pertanika Journal of Social Sciences & Humanities, 28(1), 2020.

[8] Li, C., & Liu, Y. "Joint POS tagging and text normalization for informal text". In Twenty-Fourth International Joint Conference on Artificial Intelligence, June 2015.

[9] Tufis, D., & Mason, O. "Tagging Romanian texts: a case study for qtag, a language-independent probabilistic tagger". In Proceedings of the First International Conference on Language Resources and Evaluation (LREC) (Vol. 1, No. 589-596, p. 143), May 1998.

[10] Othman, A., & Karim, N. S. "Kamus komprehensif bahasa Melayu". Penerbit Fajar Bakti, 2005.

[11] Le, T. A., Moeljadi, D., Miura, Y., & Ohkuma, T. "Sentiment analysis for low resource languages: A study on informal Indonesian tweets". In Proceedings of the 12th Workshop on Asian Language Resources (ALR12) (pp. 123-131), December 2016.

[12] Jamali, N. "Fenomena Penggunaan Bahasa Slanga dalam Kalangan Remaja Felda di Gugusan Felda Taib Andak: Suatu Tinjauan Sosiolinguistik". Jurnal Wacana Sarjana, 2(3), 1-1, 2018.

[13] Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data". In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 198-206), 2013.

[14] Nakagawa, T., Kudoh, T. & Matsumoto, Y. "Unknown word Guessing and Part-of-Speech Tagging Using Support Vector" Mac, Hines. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pp. 325—331, 2001.

[15] Zhu, B., Tokuno, J., & Nakagawa, M. "Segmentation of online handwritten Japanese text using SVM for improving text recognition". In International Workshop on Document Analysis Systems (pp. 208-219). Springer, Berlin, Heidelberg, February 2006.

[16] Giménez, J., & Marquez, L. "Fast and accurate part of speech tagging: The SVM approach revisited". Recent Advances in Natural Language Processing III, 153-162, 2004.

[17] Lee, Y. K., & Ng, H. T. "An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation". In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 41-48). Association for Computational Linguistics, July 2002.

[18] Schmid, H. "Part-of-speech tagging with neural networks". In Proceedings of the 15th conference on Computational Linguistics-Volume 1 (pp. 172-176). Association for Computational Linguistics, August 1994.

[19] Abdulkareem, M., & Tiun, S. "COMPARATIVE ANALYSIS OF ML POS ON ARABIC TWEETS". Journal of Theoretical & Applied Information Technology, 95(2), 2017.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. "Scikit-learn: Machine learning in Python". The Journal of Machine Learning Research, 12, 2825-2830, 2011.

[21] Bird, S., Klein, E., & Loper, E. "Natural language processing with Python: analyzing text with the natural language toolkit". "O'Reilly Media, Inc.", 2009.

[22] Kulkarni, A., & Shivananda, A. "Natural language processing recipes". Apress, 2019.

[23] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. "Part of speech tagging for twitter: Annotation, features, and experiments". Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011.

[24] Nooralahzadeh, F., Brun, C., & Roux, C. "Part of speech tagging for french social media data". In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1764-1772), August 2014.

[25] Màrquez, L., & Rodríguez, H. "Part-of-speech tagging using decision trees". In European Conference on Machine Learning (pp. 25-36). Springer, Berlin, Heidelberg, April 1998.

[26] Márquez, L. "Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees". Universitat Politècnica de Catalunya, 1999.