Rule-based Text Normalization for Malay Social Media Texts

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

***Abstract*—Malay social media text is a text written on social media networks like Twitter. Commonly, this text comprises non-standard words, filled with dialects, foreign languages, word abbreviations, grammatical neglect, spelling errors, and many more. It is well known that this type of text is difficult to process due to its high noise and distinct text structure. Such problems can be resolved using rigorous text normalization, which is critical before any technique can be implemented and evaluated on social media text. In this paper, an improved normalization method towards Malay social media text was proposed by converting non-standard Malay words using a rule-based model. The method normalizes common language words often used by Malaysian users, such as non-standard Malay (like dialect and slangs), Romanized Arabic, and English words. Thus, a Malay text normalizer was proposed using a set of rules that extend across different domains of natural language processing (NLP) and is expected to address the challenges of processing Malay social media text. This study implements the proposed Malay text normalizer in a Part-of-Speech (POS) tagging application to evaluate the normalizer's performance. The implementation demonstrates a substantial improvement in the POS tagging efficiency over several pre-processing stages, with an improvement of accuracy up to 31.8%. The increase of accuracy in the POS tagging indicates two main points. First, the Malay text normalizer's rules improve the performance of a Malay text normalizer on social media text. Second, our proposed Malay text normalizer has successfully improved the POS tagging percentage and demonstrates the importance of normalized pre-processing in any NLP application.**

Keywords*—*Malay normalization; Malay text normalization; informal Malay text; Malay tweets; rule-based normalizer

I. INTRODUCTION

Twitter is among the most influential social networks globally after Facebook, and it is expected to remain a common choice for years to come [1]. Twitter is a micro-blogging and social networking service that enables registered users to write, read, and share a short text with other users. A short text, called tweet writing, is limited to only 280 letters [2][3]. This restriction leads users to engage more creatively using a non-standard way of writing. For example, slang, abbreviations, emoticons, and shortcuts are [4] used to satisfy the criteria for a limited number of characters permitted by tweets. According to the study published by [1], from the first quarter of 2010 to the fourth quarter of 2018, there were 1,318 million users worldwide active monthly on Twitter accounts. These statistics clearly show that Twitter has an extensive social media text database (or texts written in colloquial or non-standard language).

Studies on the normalization of Malay social media texts are still lacking and needs improvement from various aspects. Besides, current normalization techniques are unable to normalize nearly the entire vocabulary of Malay social media content. One way to improve the quality of language processing on social media data is by using normalization methods that can automatically convert non-standard terms to its corresponding standard token [5]. Therefore, the goals of this study are to propose an improved normalization technique that can normalize Malay social media texts by converting and mapping non-standard Malay words to its corresponding standard Malay word form. This technique includes converting shorten and typographical words of Romanized Arabic and English words. The method helps to boost the efficiency of any NLP applications such as Machine Translation, Named Entity Recognition, POS Tagging, and more [6].

Designing an automated Malay text normalizer without human intervention seems to be a challenging job. Therefore, this study decides to develop a Malay text normalizer based on a set of rules, as mentioned in Section III. The critical challenge in developing text normalization for Malay social media text is building a repository that contains all non-standard word variations present in a training corpus. The repository was built from an extensive collection of Malay tweets. This study extracts the Malay tweets data manually from the Malaysian Twitter account like [9]. The data extracted are carefully selected to only tweets written using non-standard Malay language to ensure that the data used is appropriate and aligned with the objective of this study. However, due to time restriction, this study only managed to collect 1,791 tweets, and the total number of words in the data collection is only 38,714 words. The built-in repository includes 2,848 non-standard word variants that are mapped into 1,292 standard forms [7]. The most massive non-standard word variants in this category refer to the term 'beritahu' with 16 variants.

The performance of this proposed Malay text normalizer is evaluated based on an application-based method. This study adopts the approach of evaluating a text normalizer like [8]. The author in [8] evaluates a text normalization technique of English text based on the performance of a spell checkers. Whereas, for the Malay text normalizer, the performance of the POS tagging is used instead. POS is the method of tagging a word with its corresponding tags. This POS tagging

application's performance is evaluated based on the same evaluation method of [9].

This study presents the first attempt to produce a text normalizer for Malay social media content. The technique proposed consists of three steps: (i) pre-process tweets, (ii) detect non-standard words, and (iii) correct it using Malay text normalizer. This paper's organization shall be as follows: Section II presents the related works; Section III presents the method; Section IV presents the evaluation and results; and finally, the conclusion present in Section V.

II. Related Work

Social media text is challenging to read since much of the content in this area appears in dialects, slangs, abbreviations, and foreign languages. This study utilized a corpus of data from Malay's Twitter users in Malaysia. The data in this corpus must be pre-processed to remove any unnecessary sign or language mistake. The pre-processing phase is a critical step that must be completed by each researcher to be able to process their work computationally. According to [10], their study findings revealed that 17 pre-processing techniques were used to process text data from the Malay language. These techniques of pre-processing include capitalization, tokenization, spelling correction, and more. According to [10], the Sabah dialect study by [11] is the Malay text-based analysis using the most pre-processing technique. The authors in [11]'s used 7 of the 17 pre-processing techniques identified by [10]. Meanwhile, studies by [12] and [13] were using the entire six pre-processing techniques. Both studies of [12] and [13] use the same pre-processing techniques, such as stop words, tokenization, spelling correction, punctuation removal, and non-word removal.

Additionally, the studies conducted by [14], [15], [16], and [17] are using only four pre-processing techniques [10]. The authors in [14]'s used pre-processing methods such as capitalization, spam elimination, emoticon elimination, and symbol removal. [15]'s analysis used capitalization methods, selecting only tweets written or containing the language of analysis, filtering tweets that did not suit the report's features, and selecting only tweets that could be encoded and decoded. Besides, the study by [16] used pre-processing techniques such as the removal of stop words, diacritical removal, repeated removal of letters, and the removal of specific markers from social media. Lastly, [17] research used techniques for stop word elimination, tokenization, elimination of punctuation, and removal of multiple phrases.

Nevertheless, according to [18], pre-processing of data can only be achieved using eight basic techniques: (i) transforming upper case letters, (ii) eliminating punctuation, (iii) deleting words, (iv) neutralizing text, (v) correcting spelling errors, (vi) tokenizing, (vii) stemming, and (viii) lemmatizing. Furthermore, a study of social media text was conducted by [19] using eight normalization techniques to neutralize texts containing dialects and other grammatical errors. This social media text study by [19] focused on the Indonesian language; however, this study also considers other language families, like Malay. Table I shows the overall pre-processing techniques performed by [19].

TABLE I. List of Normalization Techniques used by [19]

Steps	***Pattern***	**Example**	
		Before	***After***
Convert all tokens to lowercase	ABC → abc	Media	media
Removal of the word *-nya* and *ny*	ABC*nya* → ABC	makanannya	makanan
	ABC*ny* → ABC	kelakarny	kelakar
Separate the word *lah*, *lh* dan *la*	ABClah → ABC lah	jomlah	jom lah
	ABClh → ABC lh	dialh	dia lah
	ABCla → ABC la	lantakla	lantak la
Removal of words with hyphens or 2	ABC-ABC → ABC	bunga-bunga	bunga
	ABC2 → ABC	kucing2	kucing
	ABC^2 → ABC	$arnab^2$	arnab
Removed character repetition	AABBCC → ABC	nakkkkkk	nak
Separate words with several similar words into two groups	ABABABAB → ABAB	hehehehe	hehe
Break together two or more words which offer different meanings	ABAC → AB AC	takapa taknak	tak apa tak nak
Converts typographical terms into actual types of words	AC → ABC	byk	banyak
	ABCX → ABC	cantix	cantik

In summary, considering the pattern of pre-processing techniques set out in Table I, this research has decided to adopt, incorporate and make some improvements to all of these pre-processing techniques to suit the corpus of studies containing the Malay social media texts.

III. Method

Throughout this section, the solution to the crucial challenge of normalizing Malay social media text (see Section I) is addressed in detail in the form of a rule-based Malay text normalizer. Therefore, this study constructs a semi-autonomous rule-based model to normalize the Malay social media text effectively. This study's training corpus consists of tweets written in a non-standard Malay language and mixed language. The mixed language mentioned here is a tweet written by mixing the Malay with foreign languages like English or Romanized Arabic. The rule-based Malay text normalizer is designed to normalize only words written in Malay, Romanized Arabic, and English. Other languages will not be normalized. The corpus is first converted into a lower case to reduce the scale of the vocabulary. Normalization procedures and their associated standardized forms are implemented sequentially on the vocabulary list.

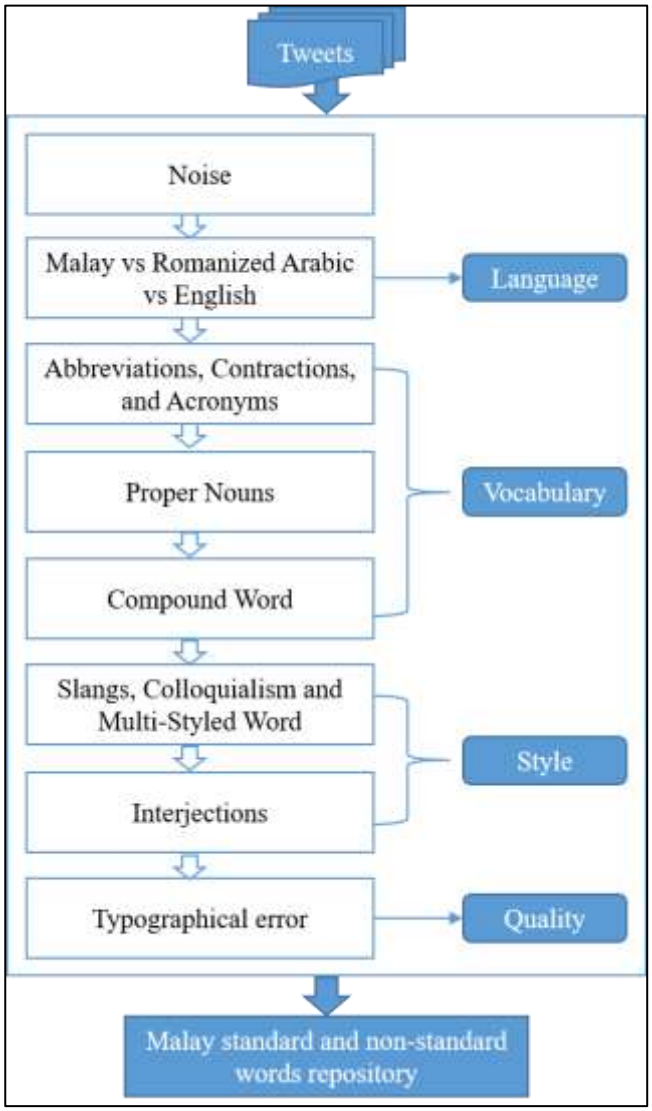

Fig. 1. The Architecture of the Rule-based Malay Text Normalizer.

This study first describes the Malay social media text's challenges and then offers solutions in normalization procedures to these challenges. The challenges are set out, and the rules used to resolve them are listed accordingly to avoid contradictory rules. Fig. 1 demonstrates the Malay text normalizer architecture based on rules that address sequentially specific problems. The rules are constructed while considering the complexities of these difficulties, thereby providing an almost ideal translation of non-standard Malay text into its structured forms.

A. Noise

Noise decides by including all the following in the document: marks, punctuations, or numerals.

1) Any word in the corpus present in Unicode is dropped. Unicode word is a term written using characters in Chinese, Korean, and Japanese.

2) Any web links are omitted from the corpus. Commonly, web links start with HTTP(s) or www.

3) Any link to the user profile that began with the '@' symbol followed by the username will be removed.

4) All punctuation characters, symbols, and numerals are omitted from the corpus.

5) All the standalone letters, multiple white spaces, and white space found in the corpus at the beginning and end sentences are eliminated.

B. Malay vs Romanized Arabic vs English

This study defines terms in Malay, Romanized Arabic, and English. It uses the rules set out in this subsection to pre-normalize terms that correspond to each language.

1) Referring to the source provided by [20] and [21], all non-standard Malay words are replaced with their standard word form.

2) Non-standard Romanized Arabic words are converted by comparison to the source provided by [20] and [21] into their correct standard form.

3) American English is used for standardizing all English words by referring to the [22] sources.

C. Abbreviations, Contractions, and Acronyms

An abbreviation is a shorter type of a sentence or phrase. A contraction is an abbreviation simplified by a word or phrase omitting the internal letters. In contrast, an acronym is an abbreviation formed from the original components of a word or phrase. All abbreviations and subclasses of these are extended to their full forms. The letter 'x' in the corpus, for example, is generally referred to as 'tak'. The letter 'x' is then transformed into its full form 'tak' or written as 'tidak'. The initial letter for Malaysia's currency, 'rm' is also extended to 'ringgit malaysia'.

D. Proper Nouns

According to [23], a proper noun is a noun that designates a being or object, does not take a restrictive modifier, and is usually capitalized. For example, 'Idris' (person), 'Johor' (place), 'Mercy Malaysia' (organization), 'Cat' (animal), and many others. All proper nouns are transformed into lowercase and standard form, as are the other words in the text like 'tranung', 'teganu', and 'ganu' to 'terengganu'.

E. Compound Word

According to [24], a compound word comprises components that are terms or any of the various combinations of words, combining forms or affixes. All compound words will be separated, normalized, or removed to simplify the next computational process. The following transformations occur for the standardization of compound words:

1) Separate words that repeat the same two letters in different groups of words like 'hehehehe' to 'hehe'.

2) Separate, normalize, and eliminate any '-nya' particles from words. For example, 'makanannya' to 'makanan'.

3) Normalize the word '-la/-lah' and distinguish it from the word. For example, 'jomlah' to 'jom lah'.

4) Remove the hyphen (-) and the word after it, and the number 2 from the word. For instance, 'bunga-bunga' to 'bunga', and 'kucing2' to 'kucing'.

5) Separate words combine several words of different meanings into different word groups like ‘taknak’ to ‘tak nak’.

F. Slangs, Colloquialism and Multi-Styled Word

A colloquialism is a word or term used in informal language. In contrast, slang is a colloquialism that is not considered typical in a language but appropriate when used in a social context.

1) Slang is normalized by the rules of the language to which it belongs. This study considered slang to be part of the vocabulary of their respective language.

2) In Malay, colloquialism used for family relationships is translated into their respective formal names. For instance, ‘kak’, ‘akak’, ‘akok’, and ‘kakok’ turn all into ‘kakak’ or ‘mok’ and ‘omak’ to ‘mak’.

G. Interjections

An interjection or exclamation is a word or expression used in a sentence to express an emotion, a feeling, or a pause. An interjection term widely used in social media text in Malaysia is the term ‘haha’. The word ‘haha’ in the interjection is a laughing phrase and found 273 times in the corpus. An interjection word that occurs in Malay social media text is usually derived from the user’s word. The word interjection must, therefore, transformed into a simple type of word. For standardizing interjections, the following transformations occur:

1) Malay expression like ‘adoh’, ‘adoi’, ‘adui’, ‘haduh’, and ‘adeh’ to ‘aduh’.

2) English words such as ‘tq’, ‘thank’, ‘thanks’, and ‘tenkiu’ to ‘thank you’.

H. Typographical Error

The typographical error described in this paper is a typing error, such as the misspelled word [25] that most likely occurred due to finger or hand slips. Any typographical mistake or sometimes shortened to typo will be corrected using a standard list of Malay spell checks.

Words that are not transformed are left unnormalized according to the rule-based mentioned in this section. It reflects the development of this text normalizer for Malay social media text. In the next section, this Malay text normalizer’s performance is evaluated by observing its efficiency on the POS tagging application. Previous studies on the POS tagging application show that utilizing text normalizer before the tagging process can yield a significant increase in the efficiency of POS tagging [7][26][5]. This increase in the performance of the POS tagging application is due to the reduction of unknown words [7], and non-standard words [5] proportion exist in the corpus after applying text normalizer [26]. Applying text normalizer on a corpus significantly impacts POS tagging performance because one of its essential features is context information [5]. A wrongly tagged word may influence the surrounding term and negatively impact the POS tagging application.

IV. Evaluation and Results

This study uses a collection of Malay tweets and domain-dependent information such as the one in the standardization repository to evaluate the proposed Malay text normalizer by applying it on POS tagging. Besides, this study also uses the research results from [9] as a benchmark to assess the proposed Malay text normalizer. In [9], they only use necessary pre-processing procedures and yet still achieves a good of POS tagging accuracy. However, by applying more efficient pre-processing, including an efficient Malay text normalizer, the POS tagging better performance is expected. Therefore, this study will attempt to use a new collection of raw Malay tweets called test corpus on several pre-processing stages settings and evaluated the results by observing its performance on the POS tagging application. All text in the test corpus is initially converted to lowercase. Subsequently, the following pre-processing procedure takes place: (1) removal of punctuations; (2) removal of symbols; (3) removal of numerals; (4) lower casing; and (5) normalization of text using Malay text normalizer. Each test corpus from various pre-processing settings will be tagged using the Malay POS tag set by an annotator.

Note that each pre-processing procedure usually reduces the resulting vocabulary (the number of unique words) in the corpus. In [9]’s study, they have used a new collection of Malay tweets as their test corpus. They used this test corpus to evaluate the performance of the POS tagging application. The test corpus was duplicated into two sets of test corpus with the same array of tweets: one set consisting of raw tweets (2,112 tokens) and the other set containing normalized tweets (2,089 tokens). The difference in total token for both sets of test corpus shows that the pre-processing procedure will reduce the corpus’s token size. Fig. 2 presents the results from [9]’s study. To be noted, the work of [9] will be the benchmark of our study.

This study assesses the efficiency of POS tagging using the same method as [9]. This study also uses the original Malay tweet collection of [9] as the test corpus.

Like [9], our proposed Malay text normalizer’s performance was evaluated on the application of POS tagging, following several steps. The first step is by tagging the data in the training corpus manually with its corresponding POS tags by an annotator who is fluent in the language present in the corpus. The POS tags used by the annotator to tag the words are Malay POS tag sets that are designed, particularly for Malay social media text. The tagged (or annotated) training corpus is then used to train the machine learning classifier to determine the POS of the words (or tokens) in the test corpus. The test corpus (from [9]’s work) is duplicated into several sets of test corpus with the same collection of tweets [9][6]. One set of test corpus is kept unnormalized (original array of tweets), and the other test corpus is prepared according to several different pre-processing procedures. These pre-processing are concerning punctuation, symbol, numeral, case, and normalization by text normalizer (see Fig. 3). After that, once again, the annotator will manually test the prediction outcome by matching the POS tags of each term in the test corpus expected by the classifier with its real POS tags of [21].

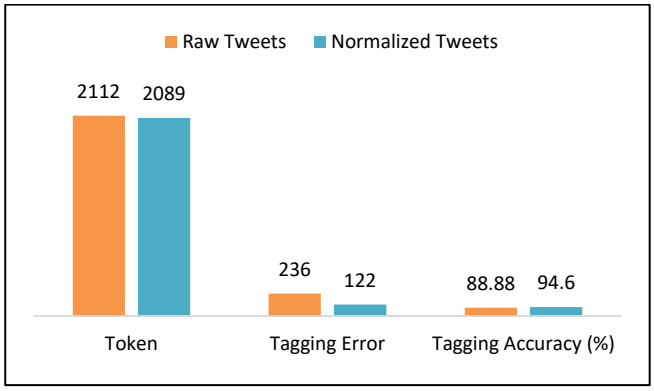

Fig. 2. The Results of [9]'s Malay Text Normalization Applied to the Malay POS Tagging.

The POS tagging efficiency of the test corpus is measured by comparing the number of tokens in the corpus, the number of tokens wrongly tagged by the classifier, and the percentage of POS tagging accuracy. The analysis even adopts and modifies [6]'s methods of assessing the NLP's performance by noting distinct changes in accuracy and the number of unique words in the various pre-processing procedures.

A. *Experimental Results*

Fig. 3 shows the POS tagging performance of the test corpus in different pre-processing procedures settings [26]; (i) comparing the token number, (ii) tagging error number, and (iii) tagging accuracy percentage. In Fig. 3, essential pre-processing procedures that seem to improve the POS tagging performance are removing punctuation, lower casing, and normalization (the process by Malay text normalizer). The use of the Malay text normalizer produces the highest impact among all pre-processing by enhancing the performance of POS tagging by up to 97.43%, thereby establishing the importance of normalization as a pre-processing process for the Malay social media text.

Note that this study used the findings of [9] (refer Fig. 2) as the benchmark of this study's POS tagging performance evaluation. The author in [9] only compared the POS tagging results between their raw and normalized test corpus. Despite using only basic normalization techniques on the normalized test corpus, [9] managed to get a high percentage in POS tagging accuracy with 94.6% (see Fig. 5). Meanwhile, even though this study used the same collection of test corpus as [9], the total number of tokens of both studies' raw tweets is different (see Fig. 4). Fig. 4 shows that this study has more token in the raw tweets test corpus compared with [9]'s study. The difference in the total token is due to the POS tagger format. QTAG POS tagger used by [9] has some restrictions on the symbols or punctuation allowed in the corpus, which [9] did not mention it clearly in their study. This study's machine learning tagger did not have any restriction, which leads to more token exist in the corpus.

The POS tagging performance of the raw tweets test corpus between these two studies (see Fig. 4) shows that [9] still managed to achieve POS tagging accuracy above 80% even with a lower total token, the lowest tagging error compared to this study. Meanwhile, in Fig. 5, the results show that this study achieved the highest POS tagging accuracy of over 97% compared to the [9]'s study and that the tagging error in this study was also the lowest. This comparison clearly shows that the proposed rule-based Malay text normalizer can reduce the proportion of unknown words. 'Unknown word' is a word that does not exist in the repository or the training corpus that causes tagging error in the test corpus and better impacts the POS tagging.

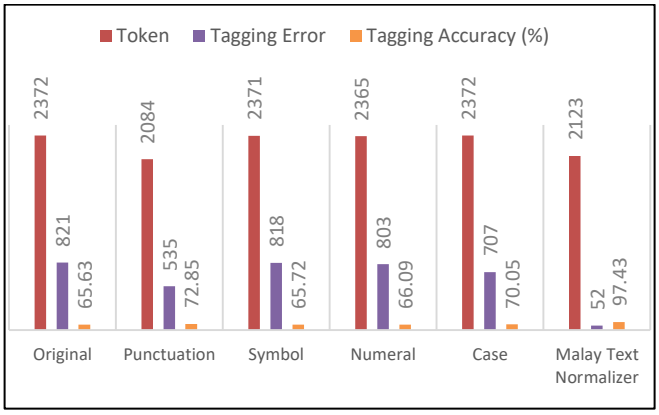

Fig. 3. POS Tagging Application Performance at different Pre-Processing Procedures.

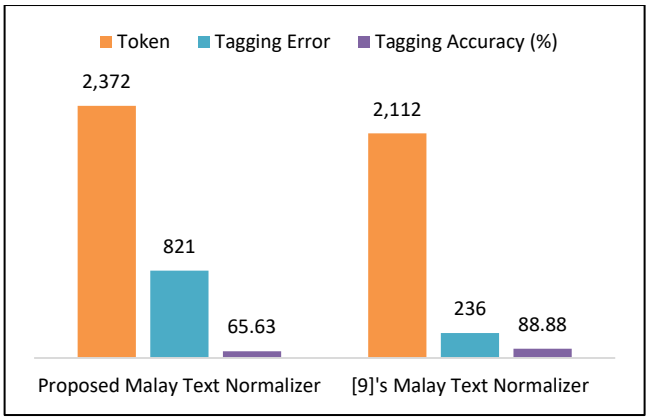

Fig. 4. The POS Tagging Performance of the Proposed Malay Text Normalizer and [9]'s Malay Text Normalizer on Raw Tweets Test Corpus.

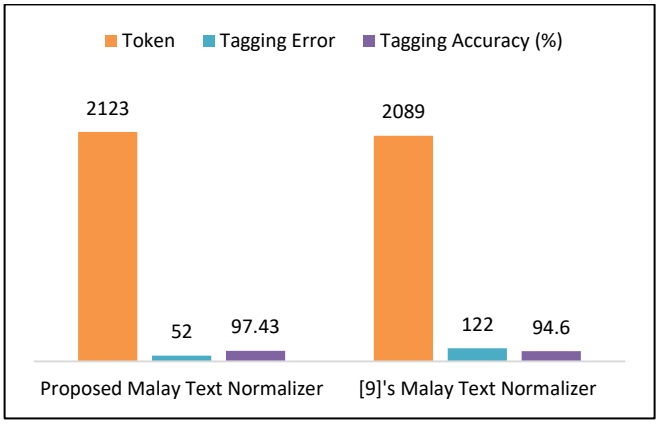

Fig. 5. The POS Tagging Performance of the Proposed Malay Text Normalizer and [9]'s Malay Text Normalizer on Normalized Tweets Test Corpus.

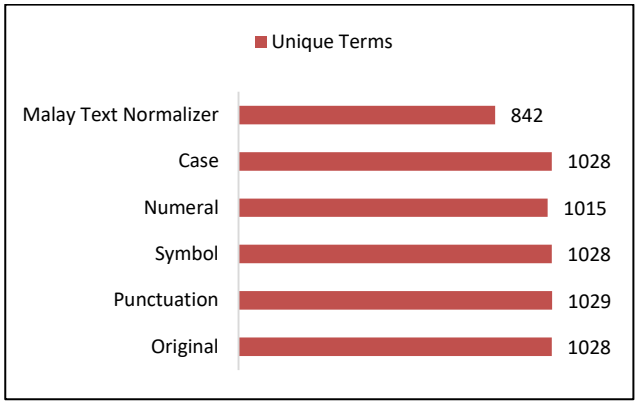

Fig. 6. The Number of unique Terms at different Pre-Processing Procedures.

This study also presents the effect of different pre-processing stages on the test corpus' vocabulary capacity. Initially, the test corpus's vocabulary size is 1,028 terms, which is reduced to 842 terms at the last step. Fig. 6 shows the vocabulary size of the test corpus after each pre-processing procedure. Significant reductions in vocabulary size are observed after applying the Malay text normalizer with a significant reduction of 18% from the original size.

Interestingly, the observed improvements in POS tagging performance are closely associated with reducing unique terms at each pre-processing level. In summary, this study concludes that the normalization of Malay social media text can yield good results in noise reduction, text normalization, and reduction of a unique term, all of which impact enhancing POS tagging efficiency.

V. Conclusion

This study aims to propose an improved text normalizer model, particularly for Malay's social media text. The proposed text normalization model converts and maps non-standard words to their corresponding standard word form. This model includes converting the simplified (or abbreviated) and typographical words of Malay, Romanized Arabic, and English.

The text normalizer for Malay social media text was built based on a rule-based approach. In which a repository that stores all non-standard word variants in the corpus was constructed. Besides, an extensive collection of Malay tweets was used as a training corpus consists of 1,791 tweets with 38,714 terms. For assessing the rule-based Malay text normalizer model, an application-based evaluation approach was used. In other words, the proposed Malay text normalizer was applied to POS tagging on social media text. The performance of the POS tagging was used to assess the performance of the proposed Malay text normalizer. In other words, if the POS tagging has improved, it also means that the proposed Malay text normalizer has improved as well.

In this study, experiments on the impact of various pre-processing stages on social media were also conducted: pre-processing punctuation, symbol, numeral, and case. The pre-processing stages with the lowest total number of the unique terms are the best pre-processing stage, and it is recommended to be used to normalize any social media text written in the Malay language. Thus, from the experiment, the text normalization by the proposed Malay text normalizer is one of the three best pre-processing (the other two are punctuation removal and lower casing).

In general, this study focuses on developing improved normalization techniques based on the modification, addition, and integration of several normalization techniques by previous studies. This study evaluated the Malay tweets collection POS tagging performance by referring to the method used by [9]. Additionally, this POS tagging efficiency can also be measured using the Hidden Markov Model, Maximum Entropy, and Support Vector Machine [27]. Besides, this research also aims to contribute to other language processing modules concentrating on the Malay social media domain, including non-standard words such as the text-to-speech system [5]. This study can be further improved by adding more non-standard Malay tweets collection into the training corpus. We assumed that by increasing the training corpus' size, the POS tagging performance for the proposed Malay text normalizer could also be increased (refer Fig. 5). The tremendous impact of our enhanced rule-based text normalizer model on the performance of an NLP application (POS tagging) proves that it is a reliable tool to normalize Malay social media content.

Acknowledgment

Universiti Kebangsaan Malaysia partially funds this research work under the research grant code: GUP-2020-063.

References

[1] Statista. (2019) The number of social network users in Malaysia from 2017 to 2023 (in millions). Retrieved Feb 20, 2019, from https://bit.ly/2D33zzt.

[2] Meftah, S., & Semmar, N. (2018, May). 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).

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

[4] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 0975-8887.

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

[6] Khan, O. A., & Karim, A. (2012, November). A rule-based model for normalization of sms text. In 2012 IEEE 24th International Conference on Tools with Artificial Intelligence (Vol. 1, pp. 634-641). IEEE.

[7] Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). 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).

[8] Clark, E., & Araki, K. (2011). Text normalization in social media: progress, problems, and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences, 27, 2-11.

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

[10] Bakar, M. F. R. A., Idris, N., Shuib, L., & Khamis, N. (2020). Sentiment Analysis of Noisy Malay Text: State of Art, Challenges, and Future Work. IEEE Access, 8, 24687-24696.

[11] Hijazi, M. H. A., Libin, L., Alfred, R., & Coenen, F. (2016, October). Bias aware lexicon-based Sentiment Analysis of Malay dialect on social

media data: A study on the Sabah Language. In 2016 2nd International Conference on Science in Information Technology (ICSITech) (pp. 356-361). IEEE.

[12] Alsaffar, A., & Omar, N. (2014, November). Study on feature selection and machine learning algorithms for Malay sentiment classification. In Proceedings of the 6th International Conference on Information Technology and Multimedia (pp. 270-275). IEEE.5.

[13] Eshak, M. I., Ahmad, R., & Sarlan, A. (2017, November). A preliminary study on hybrid sentiment model for customer purchase intention analysis in social commerce. In 2017 IEEE conference on big data and analytics (ICBDA) (pp. 61-66). IEEE.

[14] Shamsudin, N. F., Basiron, H., Saaya, Z., Rahman, A. F. N. A., Zakaria, M. H., & Hassim, N. (2015). Sentiment classification of unstructured data using lexically based techniques. Jurnal Teknologi, 77(18).

[15] Tan, Y. F., Lam, H. S., Azlan, A., & Soo, W. K. (2016, April). Sentiment Analysis for Telco Popularity on Twitter Big Data Using a Novel Malaysian Dictionary. In ICADIWT (pp. 112-125).

[16] Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M., & Omar, N. (2015). Feature selection methods effects on machine learning approach in Malay sentiment analysis. In Proc. 1st ICRIL-Int. Conf. Inno. Sci. Technol. (IICIST) (pp. 1-2).

[17] Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., & Al-bared, M. (2018). Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. PloS one, 13(4).

[18] Kulkarni, A., & Shivananda, A. (2019). Natural language processing recipes. Apress.

[19] Le, T. A., Moeljadi, D., Miura, Y., & Ohkuma, T. (2016, December). 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).

[20] Dewan Bahasa dan Pustaka. (2005). Kamus Dewan (Edisi keempat.). Kuala Lumpur: Dewan Bahasa dan Pustaka.

[21] Othman, A., & Karim, N. S. (2005). Kamus komprehensif bahasa Melayu. Penerbit Fajar Bakti.

[22] Hornby, A. S., & Omar, A. H. (2007). Oxford Compact Advanced Learner's English-Malay Dictionary. Oxford Fajar.

[23] Merriam-Webster. (n.d.). Proper noun. In Merriam-Webster.com dictionary. Retrieved August 3, 2020, from https://www.merriam-webster.com/dictionary/proper%20noun.

[24] Merriam-Webster. (n.d.). Compound. In Merriam-Webster.com dictionary. Retrieved August 3, 2020, from https://www.merriam-webster.com/dictionary/compound.

[25] Merriam-Webster. (n.d.). Typographical error. In Merriam-Webster.com dictionary. Retrieved July 31, 2020, from https://www.merriam-webster.com/dictionary/typographical%20error.

[26] Van der Goot, R., Plank, B., & Nissim, M. (2017). To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. arXiv preprint arXiv:1707.05116.

[27] Mohamed, H., Omar, N., & Juzaiddin, M. (2011). Malay Part of Speech Tagger, A comparative study on Tagging Tools. Asia-Pacific Journal of Information Technology and Multimedia. Vol, 4(1), 11-23.