ENHANCED TEXT STEMMER FOR STANDARD AND NON-STANDARD WORD PATTERNS IN MALAY TEXTS

MOHAMAD NIZAM KASSIM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

FEBRUARY 2020

DEDICATION

To The Mighty God, Most Gracious, Most Merciful.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Aizaini bin Maarof, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisors Dr. Anazida binti Zainal and Dato' Ts. Dr. Haji Amirudin bin Abdul Wahab for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented in the final copy.

My fellow postgraduate students should also be recognised for their supports. My sincere appreciation is also extended to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family members.

ABSTRACT

Text stemming is a useful language preprocessing tool in the field of information retrieval, text classification and natural language processing. A text stemmer is a computer program that removes affixes, clitics and particles to obtain the root words from the derived words. Over the past few years, few text stemmers have been developed for the Malay language but unfortunately, these text stemmers suffer from various stemming errors. It is due to the difficulty in dealing with the complexity of the Malay language morphological rules. These text stemmers are developed for text stemming against affixation words only whereas there are other affixation, reduplication and compounding words in the Malay language. Furthermore, none of these text stemmers has been developed for text stemming against social media texts which comprise of the non-standard derived words. Therefore, this research study aims to improve the existing text stemmers capability of stemming affixation, reduplication and compounding words while minimising the possible stemming errors. Moreover, this research study also aims to address text stemming process for non-standard derived words on the social media platforms by removing non-standard affixes, clitics and particles. This research study adopts a multiple text stemming approach that use affix removal method and dictionary lookup in specific arrangement order to correctly stem standard and non-standard affixation, reduplication and compounding words in the standard texts and social media texts. The proposed text stemmer is evaluated against various text documents using the direct evaluation method and the text classification is used as the indirect evaluation method to validate the effectiveness of the proposed enhanced text stemmer. In general, the proposed enhanced text stemmer outperforms the baseline text stemmer. The stemming accuracy of the proposed enhanced text stemmer achieves an average of 98.7% against the standard texts and an average of 73.7% against the social media texts. Meanwhile, the performance of the proposed enhanced text stemmer in the sports news classification application achieves an average of 85% accuracy and the illicit content classification application achieves an average of 75% accuracy. Meanwhile, the baseline text stemmer achieves an average of 63.5% stemming accuracy against the standard texts but unfortunately, it is unable to stem non-standard derived words in the social media texts. The baseline text stemmer performs poorly in sports news classification and illicit content classification with an average accuracy of 78% and 63% respectively. In short, the experimental results suggest that the proposed enhanced text stemmer has promising stemming accuracy for text stemming against the standard texts and social media texts. It also influences the performance of the text classification application.

ABSTRAK

Pencantasan teks adalah alat prapemprosesan bahasa yang berguna dalam bidang dapatan semula maklumat, pengkelasan teks dan pemprosesan bahasa tabii. Pencantas teks adalah program komputer yang membuang imbuhan, klitik dan partikel untuk mendapatkan kata dasar daripada kata terbitan. Sejak beberapa tahun lepas, beberapa pencantas teks telah dibangunkan untuk bahasa Melayu namun pencantas teks ini mempunyai pelbagai kesalahan pencantasan kata. Ia adalah disebabkan oleh kesukaran dalam menangani kerumitan peraturan morfologi bahasa Melayu. Pencantas teks ini telah dibangunkan untuk pencantasan teks bagi kata imbuhan walaupun terdapat kata imbuhan yang lain, kata ganda dan kata majmuk dalam bahasa Melayu. Selain itu, tiada satu pun daripada pencantas teks yang telah dibangunkan untuk mencantas kata terbitan bagi teks dalam media sosial yang terdiri daripada perkataan yang tidak baku. Oleh itu, kajian ini bertujuan untuk menambahbaik pencantas teks sedia ada yang mampu mencantas kata imbuhan, kata ganda dan kata majmuk sementara meminimumkan kemungkinan kesilapan pencantasan teks. Selain itu, kajian ini juga bertujuan untuk menangani proses pencantasan teks bagi perkataan tidak baku yang diperoleh dari platform media sosial dengan menyingkirkan imbuhan, klitik dan partikel yang tidak baku. Kajian ini mengambil pelbagai pendekatan pencantasan teks dengan menggunakan kaedah penyingkiran imbuhan dan kaedah pencarian kamus mengikut aturan khusus untuk mencantas kata imbuhan, kata ganda dan kata majmuk yang baku dan bukan baku dalam teks baku dan teks media sosial dengan betul. Pencantas teks yang dicadangkan telah dinilai terhadap pelbagai dokumen teks dengan menggunakan kaedah penilaian langsung dan aplikasi pengkelasan teks telah digunakan sebagai kaedah penilaian tidak langsung untuk mengesahkan keberkesanan pencantas teks yang dicadangkan. Umumnya, ketepatan pencantas teks yang dicadangkan mengatasi ketepatan pencantas teks asas. Ketepatan pencantas teks yang dicadangkan mencapai purata 98.7% pencantasan teks terhadap teks baku dan purata 73.7% pencantasan teks terhadap teks media sosial. Sementara itu, prestasi pencantas teks yang dicadangkan dalam aplikasi pengkelasan berita sukan mencapai purata ketepatan 85% dan aplikasi pengkelasan kandungan terlarang mencapai purata ketepatan 75%. Sementara itu, pencantas teks asas mencapai purata 63.5% ketepatan terhadap teks baku namun tidak dapat mencantas perkataan yang bukan baku dalam teks media sosial. Pencantas teks asas mempunyai prestasi tidak memuaskan dalam pengkelasan berita sukan dan pengkelasan kandungan terlarang dengan ketepatan purata masing-masing 78% dan 63%. Secara ringkasnya, dapatan kajian menunjukkan bahawa pencantas teks yang dicadangkan memberi ketepatan yang baik terhadap teks baku dan teks media sosial. Ia juga mempengaruhi prestasi aplikasi pengkelasan teks.

TABLE OF CONTENTS

TITLE PAGE

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Since the creation of the Internet, the number of digitised texts available to internet users have increased from time to time. However, these digitised texts have grown exponentially due to the participation of internet users on the social media platforms in recent years (Kapoor *et al.*, 2018). Although large digitised texts such as machine-readable text corpora (e.g. British National Corpus, American National Corpus, Penn Treebank) have been available for many decades, user-generated contents on the social media platforms have provided significant amount of the digitised texts which led to the rise of the modern-age text corpora in the field of text analysis to research communities (Kanan *et al.*, 2019). The use of these digitised documents can be seen in numerous active fields of research studies such as information retrieval, text classification, text clustering, natural language processing and sentiment analysis (Hansen, 2018).

These research areas generally require language preprocessing tools that can handle a large number of morphological variants of word patterns in the digitised texts, especially in morphologically rich natural languages. Specific preprocessing tools are needed for analysis at lexical, morphological, syntactic and semantic levels (Alhaj *et al.*, 2019). Among various language preprocessing tools, a text stemmer is a basic but very useful language preprocessing tool to deal with a wide range of morphological variants of word patterns in the digitised texts (Singh and Gupta, 2017b).

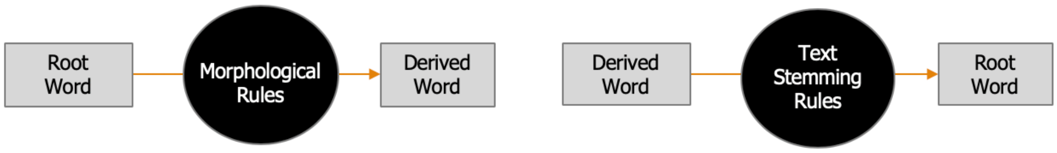
Text stemming is a process of removing bound morphemes (i.e. affixes, clitics and particles) from different morphological variants of the word patterns (derived word patterns) into their base forms (sometimes called root words or stemmed words). It is a useful language preprocessing tool to map (stem) these word patterns. The computer program used to perform text stemming process is called a text stemmer.

A text stemmer is usually developed for a specific application (Kaur and Buttar, 2018). Therefore, the research community has different perspectives concerning the use of a text stemmer in their research interests. On the first perspective, one can view a text stemmer as a compression mechanism for reducing the number of words in the text documents during the indexing process due to different morphological variants of the word patterns are mapped to a single base form. As a result, the size of the indexed text documents can be reduced to more than half the size of the original text documents (Fautsch and Savoy, 2009). On the second perspective, one can view a text stemmer as a recall/precision enhancement for improving the accuracy of text document retrieval. By mapping different morphological variants of the word patterns into their base forms, the problem of vocabulary mismatch between the user queries and the text documents is resolved during the indexing of the text documents (Kraaij and Pohlmann, 1996). Lastly, on the third perspective, one can view a text stemmer as a dimensionality reduction mechanism which reduces various morphological variants of the word patterns to their base forms which increases the likelihood of choosing appropriate features (or words) from the text documents particularly in the text applications (Boukil *et al.*, 2017). Hence, the research community uses a text stemmer as a language preprocessing tool in their fields of research (Singh and Gupta, 2017b).

Despite many text stemmers are proposed in the literature, the issue of stemming errors still remain. Stemming errors have been the core issue in the existing text stemming research, even though various text stemming approaches have been proposed to address the complexity of the morphological rules in the Malay language (Alfred *et al.*, 2014). Researchers continue to focus on improving the effectiveness of the text stemmers to map various morphological variants of the word patterns into their base forms without stemming errors (Khan *et al.*, 2017). Therefore, an enhanced text stemmer with promising stemming accuracy is highly desirable.

1.2 Background of the Problems

Text stemming research is one of the most active fields of research in many natural languages. Text stemmers have been developed for 32 different languages from 12 language families - Semitic, Turkic, Malayo-Polynesian, Balto-Slavic, Germanic, Hellenic, Indo-Iranian, Italic, Atlantic-Congo, Paman, Finnic and Ugric (Singh and Gupta, 2017b). Each of the text stemmers is developed in accordance with the morphological rules of its corresponding natural languages. From the perspective of linguistics, morphological rules are applied to transform the root words into the derived word patterns. In contrast, a text stemming rule is applied to map (stem) derived word patterns into their corresponding root words. The difference between the morphological rules and the text stemming rules is shown in Figure 1.1.



(a) Linguistic Perspective

(b) Text Stemming Perspective

Figure 1.1 Morphological Rules vs. Text Stemming Rules

The great challenge in designing and developing an enhanced text stemmer for morphologically rich natural languages is to deal with a large number of morphological variants of the word patterns. For instance, the English words *disconnect*, *disconnected*, *connects*, *connected*, *connection*, *connecting*, and *connector* are mapped to their base form *connect* with the help of a text stemmer. Due to the Malay language is one of the morphologically rich natural languages, text stemming is far more complicated than the English language. For instance, Malay words of *bersambung* (connect), *sambungan* (connection), *disambungkan* (connected to), *kesinambungan* (continuity), *bersambung-sambung* (continuously), sambung-*sinambung* (continued), *sesambung* (connector), and *ketidaksambungan* (not connected) are mapped into their base form *sambung* as shown in Figure 1.2. Thus, a text stemmer is developed for a specific natural language and cannot be used for other natural languages.

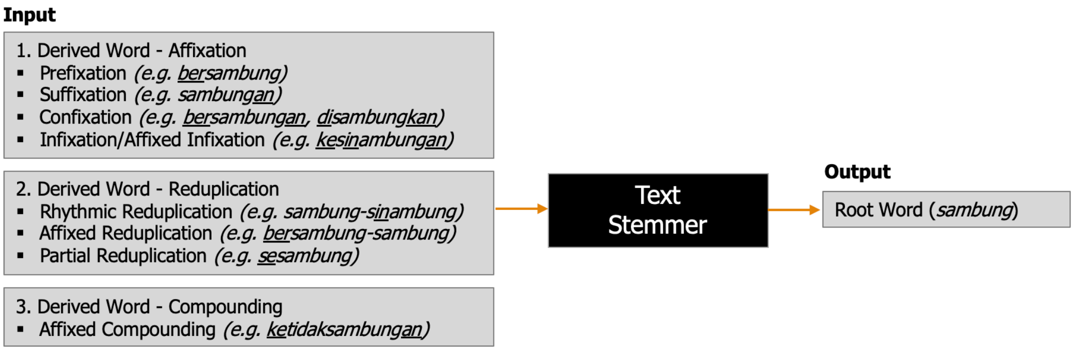

Input
1. Derived Word - Affixation
▪ Prefixation (e.g. bersambung)
▪ Suffixation (e.g. sambungan)
▪ Confixation (e.g. bersambungan, disambungkan)
▪ Infixation/Affixed Infixation (e.g. kesinambungan)
2. Derived Word - Reduplication
▪ Rhythmic Reduplication (e.g. sambung-sinambung)
▪ Affixed Reduplication (e.g. bersambung-sambung)
▪ Partial Reduplication (e.g. sesambung)
3. Derived Word - Compounding
▪ Affixed Compounding (e.g. ketidaksambungan)
Text Stemmer
Output
Root Word (sambung)

Figure 1.2 Text Stemming for the Malay Language

Each natural language has its own set of the morphological rules that may lead to their corresponding challenges in the design and development of a text stemmer. The challenges of complying morphological rules in the Malay language are quite complicated, as researchers have tried, and unfortunately, they still suffer from stemming errors (Alfred *et al.*, 2014). In order to further understand, there are seven research problems which need to be addressed, as shown in Figure 1.3.

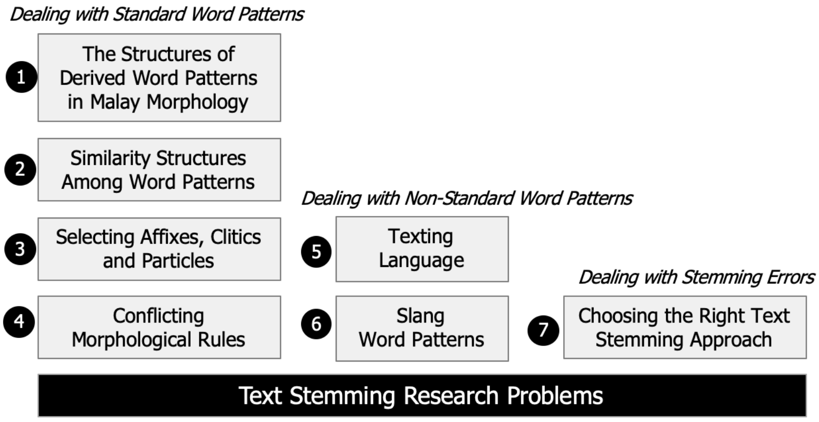
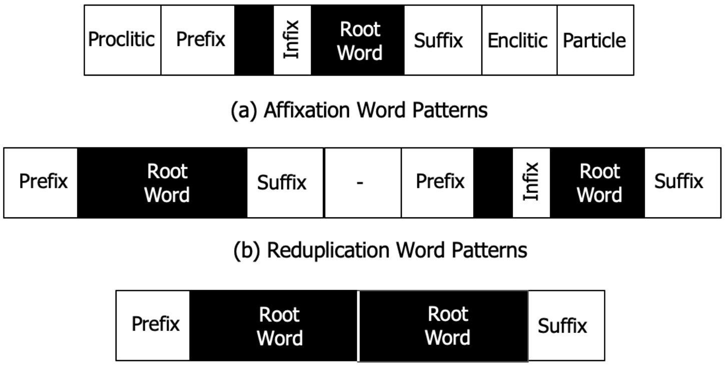

Dealing with Standard Word Patterns
1 The Structures of Derived Word Patterns in Malay Morphology
2 Similarity Structures Among Word Patterns
3 Selecting Affixes, Clitics and Particles
4 Conflicting Morphological Rules
Dealing with Non-Standard Word Patterns
5 Texting Language
6 Slang Word Patterns
Dealing with Stemming Errors
7 Choosing the Right Text Stemming Approach
Text Stemming Research Problems

Figure 1.3 Text Stemming Challenges for the Malay Language

1.2.1 The Structures of Derived Word Patterns in the Malay Morphology

According to the word morphology, there are seven word patterns in the Malay language, i.e. affixation, reduplication, compounding, blending, clipping, abbreviation and borrowing (Hassan, 1974). Only word patterns of affixation, reduplication and compounding are considered as the derived word patterns (Ranaivo-Malancon, 2004).

These word patterns have their own morphological rules on how the root words are formed into the derived word patterns by inserting affixes (prefixes, suffixes, infixes), clitics (proclitics, enclitics) and particles as shown in Figure 1.4.



(a) Affixation Word Patterns

(b) Reduplication Word Patterns

(c) Compounding Word Patterns

Figure 1.4 The Structures of Derived Word Patterns in the Malay Language

Affixation words have four subclasses, i.e. prefixation, suffixation, confixation (sometimes called circumfixes or prefix-suffix pair) and infixation words, as shown in Figure 1.4(a). These subclasses have different structures from one to another, e.g. prefixation words *bersambung* (connect), suffixation words *sambungan* (connection), confixation words *disambungkan* (connected to) and infixation words *sinambung* (continuous). The underlined syllables within the derived word patterns are called bound morphemes (i.e. affixes, clitics or particles) that are usually attached at the beginning, at the end, both at the beginning and at the end or at the middle of the root words to form affixation word patterns.

On the other hand, reduplication words also have four subclasses, i.e. full reduplication, rhythmic reduplication, affixed reduplication and partial reduplication words, as shown in Figure 1.4(b). Similar to affixation words, these subclasses also have different structures from one to another, e.g. full reduplication word *jari-jari* (fingers), rhythmic reduplication word *jari-jemari* (fingers), affixed reduplication word *jari-jarinya* (his/her fingers) and partial reduplication word *jejari* (radius). It is important to note that full reduplication, rhythmic reduplication and affixed reduplication words have a hyphen (-) between two words except partial reduplication words. These word patterns have different structures from affixation word patterns due to these words have repetition in words, hyphen (-) between the words and sometimes, these words may have bound morphemes attached at the beginning, at the end, both at the beginning and at the end, or at the middle of the root words to form affixed reduplication word patterns.

The last category of derived word patterns in the Malay language is compounding words. There are two subclasses, i.e. free-form compounding words and affixed compounding words, as shown in Figure 1.4(c). Free-form compounding words are derived from a combination of two root words such as two root words, *tanggung* (bear) and *jawab* (answer) to form derived words, *tanggungjawab* (responsibility). These word patterns are different from reduplication word patterns which these word patterns are not separated by a hyphen (-) to form compounding words. Another form of compounding words is affixed compounding words. Affixed compounding words are derived from free-form compounding words, e.g. *ambil alih* (take over) with bound morpheme attached at the beginning or both at the beginning and at the end of the free-form compounding words to form affixed compounding word patterns, e.g. *pengambilalihan* (merger and acquisition).

It can be concluded that the Malay language is a morphologically rich language where there are morphological variants of affixation, reduplication, and compounding words with their corresponding subclasses. From the perspective of text stemming research for Malay language, none of the existing text stemmer was developed for affixation, reduplication, and compounding words. Most of the existing text stemmers are developed for affixation words only (Alfred *et al.*, 2014). However, there are some existing text stemmers which are developed for affixation words and also, partial subclasses of the reduplication words (Fadzli *et al.*, 2012; Khan *et al.*, 2017). Applying affixation stemming rules against reduplication or compounding words are considered not appropriate due to these derived words have different morphological rules in the Malay language. The following scenarios are the text stemming challenges which relate to the structures of derived word patterns:

i. Removing the second part of reduplication word with a hyphen (-) and then, applying affixation stemming partially solves the problems but remain unsolved for other structures of word patterns which are similar to reduplication words, e.g. *pro-kerajaan* (pro-government), *anti-kemajuan* (anti-development), *Kedah-Kelantan* (showing between two destinations).

ii. Partial reduplication words, e.g. *jejari* (radius), *cecair* (liquid) cannot be stemmed by applying affixation stemming rules due to syllable *je, ce* are not part of affixes, clitics or particles.

iii. Applying affixation stemming rules against affixed compounding words are not sufficient to obtain correct root words, e.g. *pengambilalihan* (merger and acquisition) contains two root words *ambil* (to take) and *alih* (to move) and not *ambilalih* (to takeover).

Therefore, the key research problem that needs to be addressed is to develop text stemming rules for affixation, reduplication, and compounding words based on their corresponding morphological rules.

1.2.2 Similarity Structures Among Word Patterns

Due to the large numbers of morphological variants of word patterns, there are chances of having similar structures among word patterns. It leads to the difficulty to design and develop text stemming as the differentiation among these word patterns are crucial for eliminating various possible stemming errors. These stemming errors have been reported in the existing text stemming research. The following scenarios are text stemming challenges which relate to the similarity structures among word patterns:

i. Non-derived word patterns (root words) may have similar structures with affixation words, e.g. *belia* (youth), *makan* (eat), *diari* (diary).

ii. Non-derived word patterns (root words) may have similar structures with compounding words, e.g. *jawatankuasa* (committee).

iii. Prefixation and suffixation words may have similar structures with confixation words, e.g. *berjalan* (to walk), *bekalan* (supply), *kurangkan* (to reduce).

iv. Affixed reduplication words may have similar structures with full reduplication words, e.g. *perompak-perompak* (robbers), *makanan-makanan* (foods), *pelajaran-pelajaran* (lessons).

v. Affixed compounding words may have similar structures with confixation words, e.g. *pengambilalihan* (takeover).

From the perspective of text stemming research for the Malay language, only differentiation between root words and affixation words are considered in the existing text stemmers (Darwis *et al.*, 2012; Khan *et al.*, 2017). Therefore, the key research problem that needs to be addressed is to differentiate these word patterns so that text stemming rules are applied appropriately in accordance with morphological rules.

1.2.3 Selecting Affixes, Clitics and Particles from the Derived Word Patterns

There are three different types of bound morphemes in the Malay morphology, i.e. affixes, clitics and particles. These bound morphemes usually are attached at the beginning, at the end, both at the beginning and the end, or at the middle of the root words, to form derived word patterns as shown in Figure 1.4.

Affixes have four different subclasses, i.e. prefixes (attached at the beginning of the root words), suffixes (attached at the end of the root words), confixes (at the beginning and the end of the root words) and infixes (attached at the middle of the root words). Most common affixes are prefixes (*ber+, di+, ke+, men+, pen+, per+, se+, ter+*), suffixes (*+an, +i, +kan)* and *confixes* (*ber+an, ber+kan, di+i, di+kan, ke+an, memper+i, memper+kan, men+i, men+kan, pen+an, per+an*). On the other hand, clitics are another form of bound morphemes that are attached at the beginning of derived word patterns called proclitics (*ku+* and *kau+)* and at the end of derived word patterns called enclitics (*+ku*, *+mu* and *+nya)*. Lastly, particles are also bound morphemes that are attached at the end of the derived word patterns. These particles are *+kah, +lah, +pun* and *+tah*.

Different combinations of affixes, clitics and particles create a significant challenge in text stemming research where the order of affixes, clitics, and particles to be removed from the derived word patterns must be correctly determined. Failure to select the correct combination of affixes, clitics and particles from the derived word patterns will cause overstemming or understemming errors during the text stemming process. These stemming errors have been reported in the existing text stemming research (Leong *et al.*, 2012; Yasukawa *et al.*, 2009). The following scenarios are text stemming challenges which relate to selecting affixes, clitics and particles from the derived word patterns:

i. Overstemming errors occur when the word, e.g. *berkesan* (effective) should be stemmed as *kesan* (effect) if the correct affixes are selected instead of *san, kes*, *s* if wrong affixes are selected.
ii. Understemming errors occur when the word, e.g. *berkebolehan* (capable of) should be stemmed as *boleh* (able) if the correct affixes are selected instead of *keboleh*, *bolehan* which there are remaining affixes that must be removed.

From the perspective of text stemming research for the Malay language, the key research problem needs to correctly select and remove affixes, clitics and particles from the derived word patterns so that overstemming or understemming errors can be avoided.

1.2.4 Conflicting Morphological Rules

Malay morphology has the concept of spelling variations where prefixes *me*+ are allowed to attach with root words beginning with letters: *l, m, n, ng, ny, r* and *w*, prefixes *mem*+ and *pem*+ are allowed to attach with root words beginning with letters: *b* and prefixes *men*+ and *pen*+ are allowed to attach with root words beginning with letters: *c, d, j, t, y* and *z*. Another concept in Malay morphology is special exceptions where prefixes *mem*+ and *pem*+ will drop the first letter of the root words beginning with letters: *f* and *p*, prefixes *men*+, *pen*+ and *sepen*+ will drop the first letter of the root words beginning with letters: *t,* prefixes *meng*+ and *peng*+ will drop the first letter of the root words beginning with letters: *k*, and prefixes *meny+* and *peny+* will drop the first letter of the root words beginning with letters: *s*.

Generally, two different scenarios need to be addressed, i.e. text stemming for simple morphological rules and text stemming for conflicting morphological rules. Text stemming for simple morphological rules refers to the single process of removing affixes, clitics and particles from the derived word patterns such as *alkitab* (the book), *laluan* (path) and *bersabarlah* (be patient). Meanwhile, text stemming for conflicting morphological rules refer to the issues of spelling variations and special exceptions. As a result, the text stemmer has difficulty to stem specific structures of derived word patterns due to the issues of conflicting in spelling variation and special exception rules that may lead to overstemming or understemming errors. These stemming errors have also been reported in the existing text stemming research (Ahmad *et al.*, 1996; Idris and Mustapha, 2001; Lee *et al.*, 2013). The following scenarios are text stemming challenges which relate to conflicting morphological rules:

- Conflicting scenario to apply special exception rules, e.g. *memikir* (to think) and *memukul* (to beat), *pemikir* (thinker) and *pemukul* (beater)
- Conflicting scenario to apply spelling variation and special exception rules, e.g. *menyanyi* (to sing) and *menyapu* (to sweep), *penyanyi* (singer) and *penyapu* (sweeper)

From the perspective of text stemming research for the Malay language, the key research problem that needs to be addressed is to apply the correct spelling variations and special exceptions rules so that overstemming or understemming errors can be avoided.

1.2.5 Non-Standard Word Patterns - Texting Language

Texting language refers to the use of shortened words in online communication which is intended to reduce the number of letters to be inserted on a smartphone or computer keyboard so that it leads to a faster communication (Yeo and Ting, 2017).

The shortened words may be written in various forms on the social media platform as follows:

i. Removing all vowels of the word, e.g. *sklh* (*sekolah*), using the first syllable of the word, e.g. *sem* (*semester*), dropping the first vowel of the word, e.g. *sapa* (*siapa*) or dropping the last vowel of the word, e.g. *ank* (*anak*)

ii. Using the last syllable of the word, e.g. *mak* (*emak*) or using of the first letter and the last syllable of the word, e.g. *bleh* (*boleh*)

iii. Using the first and last letters of the word, e.g. *yg* (*yang*), using the first letters in the phrase, e.g. BNM (Bank Negara Malaysia)

iv. Using the apostrophe (‘) to drop one or two syllables, e.g. *k'jaan* (*kerajaan*)

These shortened words may be combined with affixes, clitics and particles to form derived word patterns which may be written in non-standard forms such as *ber*+(*br*+) and *pro*+ (*pro-*+). From the perspective of the text stemming research for Malay language, the problem arises in text stemming when the word, *brsklh* (*bersekolah*) should be stemmed to *sekolah* (not *brsklh* or *sklh*) and the word, *pro-k'jaan* (*pro-kerajaan*) should be stemmed to *raja* (not *prok'jaan*, *k'jaan* or *k ja*). To the best of our knowledge, none of the existing text stemmers is developed to normalise or to stem these non-standard derived word patterns (Arif and Mustapha, 2017; Samsudin *et al.*, 2012). Hence, the key research problem that needs to be addressed is to correctly normalise or stem these non-standard derived word patterns on the social media platforms.

1.2.6 Non-Standard Word Patterns - Slang Word Patterns

Other than texting language, colloquial language has also been used in the online communication. Colloquial language refers to slang words used in daily conversation and has been considered inappropriate in the official environment (Marzuki, 2013; Samsudin *et al.*, 2011; Subagyo, 2007). These word patterns may be written in various forms on the social media platforms as follows:

i. Replacing the last letter ‘a’ with ‘e’ e.g. *ape* (*apa*), *harge* (*harga*).
ii. Replacing the last two letters ‘ar’ with ‘o’ or ‘aq’ e.g. terbakaq (*terbakar*) for dialect in Northern states.
iii. Adding the letter ‘*g*' at the end, e.g. *makang* (*makan*) for Terengganu dialect.

These slang words can also be combined with affixes, clitics and particles to form derived word patterns and may be written in non-standard forms, e.g. *ber*+(*br*+), *+kan* (*+kn*), and *+lah* (*+la, +le +laa, +lar*). From the perspective of text stemming research for the Malay language, the problem arises in text stemming when the word, *punyalah* (to stress something about possession) can be written as ‘*punyelah*’, ‘*punyelaa*’, ‘*punyelar*’ and ‘*punyela*' with non-standard affixes. As a result, it is difficult to obtain the correct root words as the existing text stemmers for the Malay language do not consider these derived word patterns in their text stemming approach. Similar to the issues of texting language, the key research problem that needs to be addressed is correctly normalise or stem these non-standard derived word patterns in social media platforms.

1.2.7 Choosing the Right Text Stemming Approach

Finally, it is difficult to choose the right text stemming approach to map (or stem) a large number of morphological variants (including texting language and slang word patterns) into their corresponding base forms. It is important to note that there is no standard method in developing a text stemmer due to large morphological variants of word patterns in the Malay language (Hassan, 1974; Ranaivo-Malancon, 2004). Regardless of various text stemming approaches are proposed in the literature, the existing text stemmers still suffer from various stemming errors even though researchers only focus on text stemming for affixation words. From the perspective of text stemming research for the Malay language, the literature has reported that the root causes of these stemming errors are due to the reasons as follows:

i. Order of longest/shortest affixes, clitics and particles selection and removal (Ahmad *et al.*, 1996; Idris and Mustapha, 2001).

ii. Order of affix removal method to stem prefixation, suffixation, confixation and infixation words (Abdullah *et al.*, 2009; Idris and Mustapha, 2001; Othman, 1993).

iii. Order of dictionary lookup and the numbers of dictionary entries (Darwis *et al.*, 2012).

Therefore, it is crucial to further investigate the root causes of various possible stemming errors that can happen during the text stemming processes, not only, for affixation words but also, reduplication and compounding words. These investigations provide an understanding on design parameters for an enhanced text stemmer which will address the issues of arrangement order of longest/shortest affixes, clitics and particles selection and removal, arrangement order of text stemming rules for affixation, reduplication and compounding words, arrangement order of dictionary lookup and also, the number of dictionary entries. Therefore, there is a need to perform text stemming process correctly for standard and non-standard word patterns and to minimise possible stemming errors.

1.3 Statement of the Problem

Despite many existing literatures on text stemming research for the Malay language, there are still many open issues in improving the effectiveness of existing text stemmers for promising stemming accuracy. These open issues include a partial consideration in Malay morphology, which only focuses on affixation word patterns instead of affixation, reduplication and compounding word patterns, the existence of non-standard word patterns on the social media platforms and also, the limitation of the existing text stemming approach that lead to possible stemming errors. The research problem can be described as follows:

"The performance of the text stemmer can be improved by implementing text stemming rules for standard and non-standard affixation, reduplication and compounding word patterns in specific intra stemming rules within subclasses of derived word patterns

and inter stemming rules among derived word patterns, as well as by implementing various dictionary lookups to suppress possible stemming errors."

The main research question is as follows:

"*How to design and develop an enhanced text stemmer for the Malay language that considers all possible standard and non-standard derived word patterns in standard texts and social media texts and at the same time, suppressing the root causes of possible stemming errors?*"

Therefore, there are three different Research Hypothesis (RH) in this research study as follows:

RH1: Does the stemming accuracy of the proposed enhanced text stemmer perform better than the baseline text stemmer?

- Null Hypothesis, H_0 : *The stemming accuracy of the proposed enhanced text stemmer has no difference from the baseline text stemmer.*

- Alternative Hypothesis, H_1: *The stemming accuracy of the proposed enhanced text stemmer performs better than the baseline text stemmer.*

RH2: Does the performance of text classification application for standard texts using the proposed enhanced text stemmer better than text classification application using the baseline text stemmer?

- Null Hypothesis, H_0 : *The performance of text classification application for standard texts using the proposed enhanced text stemmer has no difference from text classification application using the baseline text stemmer.*

- Alternative Hypothesis, H_1: *The performance of text classification application for standard texts using the proposed enhanced text stemmer performs better than text classification application using the baseline text stemmer.*

RH3: Does the performance of text classification application for non-standard texts using the proposed enhanced text stemmer better than text classification application using the baseline text stemmer?

- Null Hypothesis, H_0 : *The performance of text classification application for non-standard texts using the proposed enhanced text stemmer has no difference from text classification application using the baseline text stemmer.*

- Alternative Hypothesis, H_1: *The performance of text classification application for non-standard texts using the proposed enhanced text stemmer performs better than text classification application using the baseline text stemmer.*

1.4 Purpose of Research

The purpose of this research is to design and develop a text stemming approach that is capable of text stemming against derived words in the standard texts and social media texts. The expected capabilities are word normalisation for non-standard word patterns and affixes, clitics and particles removal from the derived word patterns. Hence, the desired output of this research study is an enhanced text stemmer with specific arrangement order of affixation, reduplication and compounding stemming rules and also, robust from various possible stemming errors.

1.5 Objectives of the Research

The specific objectives of the research study are as follows:

i. To design and develop an enhanced text stemmer based on the structures of derived word patterns and their corresponding morphological rules, and at the same time, addresses possible root causes of stemming errors.

ii. To evaluate the effectiveness of the proposed enhanced text stemmer to stem against specific structures of derived word patterns into the correct root words by removing standard and non-standard affixes, clitics and particles and also, normalising non-standard word patterns.

iii. To validate the performance of the proposed enhanced text stemmer as a language preprocessing tool in text classification applications.

1.6 Scopes of the Research

The scope of this research study shall be limited to the following:

i. The design and development of the proposed enhanced text stemmer only focus on selecting and removing affixes, clitics and particles from the derived word patterns and does not consider part-of-speech of word patterns.

ii. The design and development of the proposed enhanced text stemmer only focus on stemming accuracy instead of time-based performance during stemming processes or stemming process in a big data analytics environment.

iii. Only special characters removal, case folding, and text stemming are considered as part of the core engine in the proposed enhanced text stemmer, and common words removal (stop words) is not considered in order to retain the original structure of the input texts.

iv. Direct evaluation method is used to measure the effectiveness of the proposed enhanced text stemmer i.e. measuring the stemming accuracy of the resultant words (root words) from the perspective of the morphological rules.

v. Indirect evaluation method is used to validate the performance of the proposed enhanced text stemmer from the perspective of text classification application.

vi. The proposed enhanced text stemmer considers standard texts and social media texts from accessible datasets among previous researchers, i.e. modern texts (news articles), academia texts (thesis abstracts), religious texts (chapters of the Malay translation of the Holy Qur'an) and popular social network platforms among Malaysians.

1.7 Importance of the Research

From a theoretical and practical point of view, this research study is essential and significant in text stemming research for the Malay language. Therefore, the rationale and motivation for this research are as follows:

i. There are different structures of word patterns in Malay morphology. While the existing text stemming research focuses on affixation word patterns, this research study focuses not only on affixation word patterns but also on reduplication and compounding word patterns. These word patterns exist in many different text documents such as modern texts, scientific texts and religious texts. Therefore, text stemming rules should be developed according to their corresponding morphological rules.

ii. The use of texting language and slang word patterns on the social media platform as a medium of communication leads to the existence of non-standard word patterns. These word patterns do not follow the standard morphological rules, and sometimes, these word patterns are incorrectly spelled. The existing text stemmers are not able to stem (or process) these word patterns as their text stemming rules only focuses on standard word patterns. Therefore, text stemming rules must be developed not only for standard word patterns but also for non-standard derived word patterns.

iii. Despite many text stemmers are developed for many years, stemming errors are still open issues due to the complexity of the morphological rules in the Malay language. The researchers have agreed that there is no fixed stemming approach to stem various derived word patterns. The stemming approach requires careful consideration of selecting and removing affixes, clitics and particles from the derived word patterns (shortest/longest match and removal), the arrangement of text stemming rules (order of rules) and also, word differentiation mechanism among derived word patterns (similarity structures). These consideration factors, if not considered, will produce stemming errors during the text stemming process. Therefore, text stemming rules should be developed not only for standard and non-standard derived word patterns but also with promising stemming accuracy.

For these reasons, the proposed enhanced text stemmer will benefit researchers and practitioners as one of the useful language preprocessing tools in many text applications for stemming standard and non-standard affixation, replication and compounding words in standard texts and social media texts.

1.8 Research Contributions

This section highlights the contribution of this research study, as shown in Figure 1.5.

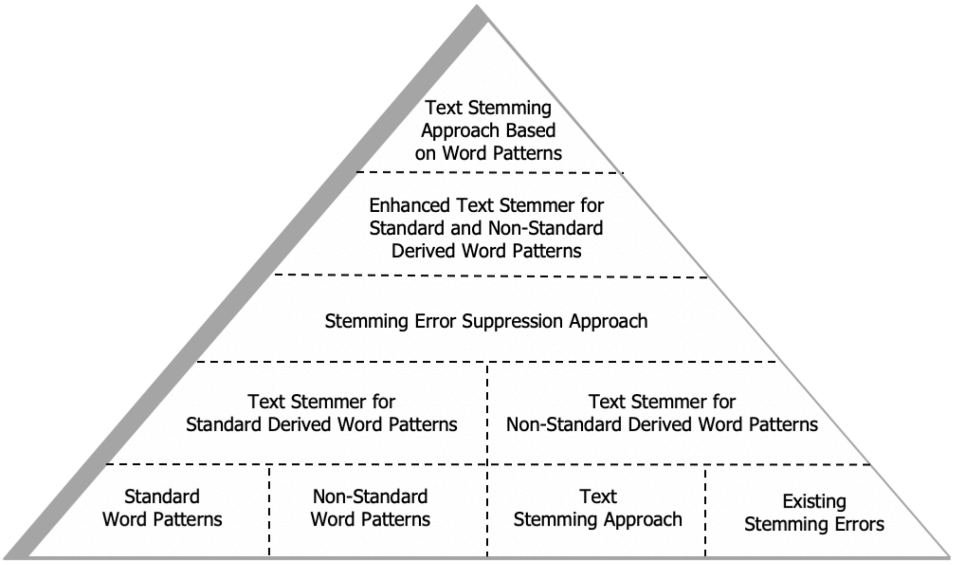

Text Stemming Approach Based on Word Patterns
Enhanced Text Stemmer for Standard and Non-Standard Derived Word Patterns
Stemming Error Suppression Approach
Text Stemmer for Standard Derived Word Patterns
Text Stemmer for Non-Standard Derived Word Patterns
Standard Word Patterns
Non-Standard Word Patterns
Text Stemming Approach
Existing Stemming Errors

Figure 1.5 Summary of Research Contributions

This research study adopts the philosophy of *'text stemming approach based on word patterns'* by introducing the specific arrangement of intra stemming rules within subclasses of the derived words and inter stemming rules among derived words. The main contribution in this research study is the proposed enhanced text stemmer for standard and non-standard derived word patterns which comprises of three supporting contributions, i.e. a text stemmer for standard derived word patterns in standard texts, a text stemmer for non-standard derived word patterns in the social media texts and a stemming error suppression approach. The consideration factors in designing and developing an enhanced text stemmer depend on standard word patterns, non-standard word patterns, text stemming approach, and the existing stemming errors. The reasoning behind these consideration factors is to address the complexity of the morphological rules for various derived word patterns and to overcome the root causes of possible stemming errors during the text stemming process.

1.9 Organisation of the Research

As shown in Figure 1.6, the thesis is structured into seven chapters. This chapter highlights open issues and challenges in the existing text stemming research that leads to the formulation of the research problem. This chapter also discusses the purpose and objectives of the research, the significance of the research and also contributions from this text stemming research.

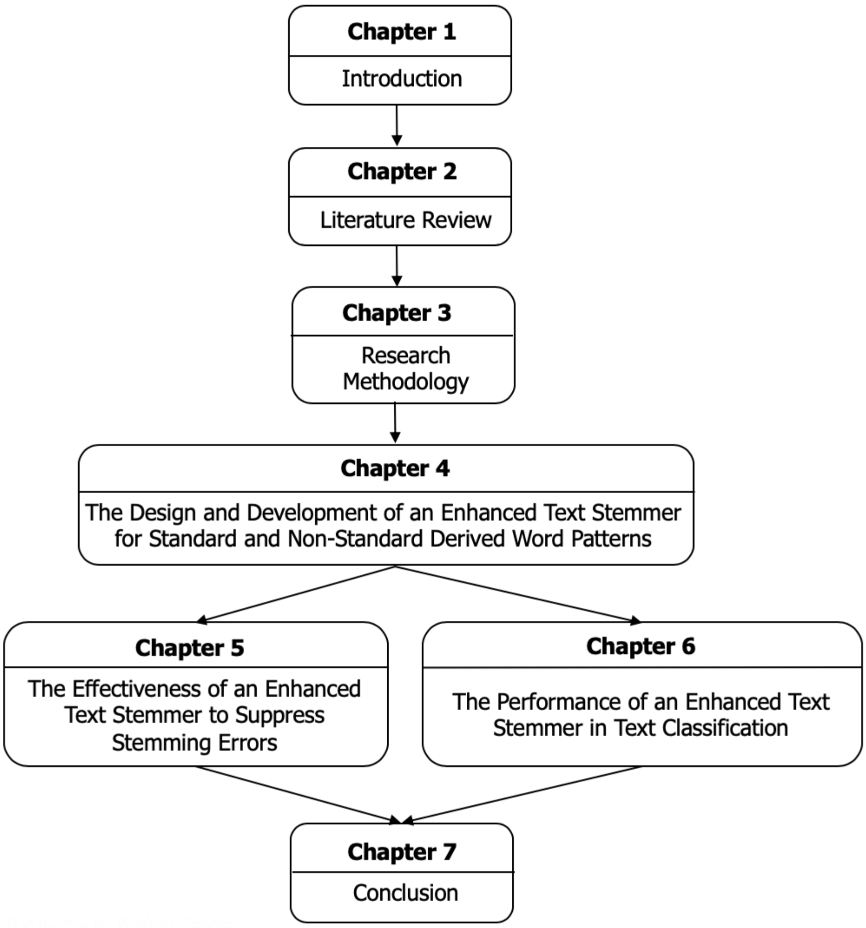

Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Research Methodology
Chapter 4
The Design and Development of an Enhanced Text Stemmer for Standard and Non-Standard Derived Word Patterns
Chapter 5
The Effectiveness of an Enhanced Text Stemmer to Suppress Stemming Errors
Chapter 6
The Performance of an Enhanced Text Stemmer in Text Classification
Chapter 7
Conclusion

Figure 1.6 Organisation of the Thesis

Chapter 2 discusses related works of existing text stemming researches involving open issues on text stemming for Malay texts. This chapter also describes morphological rules for affixation, reduplication and compounding words in the Malay language, which have been used as a basis for the design and development of the proposed enhanced text stemmer.

Chapter 3 describes the research methodology for achieving the objective of this research study. This chapter also discusses research framework, training and testing datasets, and experimental designs and procedures which have been used to develop the enhanced text stemmer which is capable of stemming standard and non-standard word patterns in standard texts and social media texts.

Chapter 4 focuses on the design and development of the proposed enhanced text stemmer. This chapter examines various text stemming approaches in order to address various stemming challenges that can lead to possible stemming errors. By understanding these challenges, design principles are introduced that will serve as the basis for developing the proposed enhanced text stemmer.

Chapter 5 focuses on the effectiveness of the proposed enhanced text stemmer to stem standard and non-standard derived word patterns. While the existing text stemming researches evaluate the effectiveness of their text stemmers against selected text documents, this research study examines the effectiveness of the proposed enhanced text stemmer against specific derived word patterns. This chapter provides an extensive evaluation of the proposed enhanced text stemmer against specific structures of standard and non-standard affixation, reduplication and compounding word patterns. The experimental results are also compared with the baseline text stemmer in relation to specific structures of derived word patterns.

Chapter 6 focuses on the performance of the proposed enhanced text stemmer as a language preprocessing tool in text classification tasks. The first experiment is conducted to evaluate the performance of the proposed enhanced text stemmer in news classification task using standard texts. Meanwhile, the second experiment is conducted to evaluate the performance of the proposed enhanced text stemmer in illicit content classification task using social media texts. The experimental results are also compared with the baseline text stemmer in relation to the performance of classification application tasks.

Finally, Chapter 7 revisits the objectives of this research, highlights the research contributions, and then concludes this research with recommendations for future research.

REFERENCES

Abdullah, M. T., Ahmad, F., Mahmod, R. and Sembok, T. M. (2009). Rules Frequency Order Stemmer for Malay Language. *IJCSNS International Journal of Computer Science and Network Security, 9*(2), 433-438.

Agusta, L. (2009). Comparison of Porter Stemming Algorithm and Nazief-Adriani's Algorithm for Stemming Indonesian Text Documents. *National Conference and Information Systems*, Yogyakarta, Indonesia: Telkom University Indonesia, 5215-5222.

Ahmad, F., Yusoff, M. and Sembok, T. M. (1996). Experiments with a Stemming Algorithm for Malay Words. *Journal of the American Society for Information Science, 47*(12), 909-918.

Alfred, R., Leong, L. C., On, C. K. and Anthony, P. (2014). A Literature Review and Discussion of Malay Rule-Based Affix Elimination Algorithms. *The 8th International Conference on Knowledge Management in Organizations*, Kaohsiung, Taiwan: Springer Netherlands, 285-297.

Alhaj, Y. A., Xiang, J., Zhao, D., Al-Qaness, M. A., Elaziz, M. A. and Dahou, A. (2019). A Study of the Effects of Stemming Strategies on Arabic Document Classification. *IEEE Access, 7*, 32664-32671.

Ariadi, D. and Fithriasari, K. (2016). Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification Dan Support Vector Machine Dengan Confix Stripping Stemmer. *Jurnal Sains dan Seni ITS, 4*(2), 248-253.

Arif, S. M. and Mustapha, M. (2017). The Effect of Noise Elimination and Stemming in Sentiment Analysis for Malay Documents. *Proceedings of the International Conference on Computing, Mathematics and Statistics (iCMS 2015)*, Langkawi, Malaysia: Springer Singapore, 93-102.

Bougar, M. (2019). Stemming Algorithm for Arabic Text Using a Parallel Data Processing. *Third International Congress on Information and Communication Technology*, Brunel University, London: Springer Singapore, 261-268.

Boukil, S., El Adnani, F., El Moutaouakkil, A. E., Cherrat, L. and Ezziyyani, M. (2017). Arabic Stemming Techniques as Feature Extraction Applied in Arabic Text Classification. *International Conference on Advanced Information*

Technology, Services and Systems, Tangier, Morocco: Springer International Publishing, 349-361.

Cohen, J. (1992). Statistical Power Analysis. *Current Direction in Psychological Science, 1*(3), 98-101.

Dansieh, S. A. (2011). Sms Texting and Its Potential Impacts on Students' Written Communication Skills. *International Journal of English Linguistics, 1*(2), 222.

Darwis, S. A., Abdullah, R. and Idris, N. (2012). Exhaustive Affix Stripping and a Malay Word Register to Solve Stemming Errors and Ambiguity Problem in Malay Stemmers. *Malaysian Journal of Computer Science, 25*(4), 196-209.

de Oliveira, R. A. and Júnior, M. C. (2017). Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence. *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017)*, Porto, Portugal: SCITEPRESS – Science and Technology Publications, 99-105.

Dutta, S., Saha, T., Banerjee, S. and Naskar, S. K. (2015). Text Normalization in Code-Mixed Social Media Text. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, Kolkata, India: IEEE, 378-382.

Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H. and Siti Dhalila, M. S. (2012). Simple Rules Malay Stemmer. *The International Conference on Informatics and Applications (ICIA 2012)*, Kuala Terengganu, Malaysia, 28-35.

Fautsch, C. and Savoy, J. (2009). Algorithmic Stemmers or Morphological Analysis? An Evaluation. *Journal of the American Society for Information Science and Technology, 60*(8), 1616-1624.

Frakes, W. B. and Fox, C. J. (2003). Strength and Similarity of Affix Removal Stemming Algorithms. *ACM SIGIR Forum, 37*(1), 26-30.

Gharatkar, S., Ingle, A., Naik, T. and Save, A. (2017). Review Preprocessing Using Data Cleaning and Stemming Technique. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India: IEEE, 1-4.

Gupta, V. and Lehal, G. S. (2013). A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages. *Journal of Emerging Technologies in Web Intelligence, 5*(2), 157-161.

Han, P., Shen, S., Wang, D. and Liu, Y. (2012). The Influence of Word Normalization in English Document Clustering. *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, Zhangjiajie, China: IEEE, 116-120.

Hansen, S. (2018). Introduction to Text Mining. *IFC Bulletin Chapters, 50*, 23-25.

Hapsari, E. D. (2018). Analisis Pengaruh Bahasa Alay (Gaul) Dalam Penulisan Pesan Melalui Sms/Wa Mahasiswa Teknik Informatika Universitas Pgri Madiun. *Linguista: Jurnal Ilmiah Bahasa, Sastra, dan Pembelajarannya, 2*(1), 29-38.

Hassan, A. (1974). *The Morphology of Malay*: Dewan Bahasa Dan Pustaka, Kementerian Pelajaran Malaysia.

Hidayatullah, A. and Ma'arif, M. (2017). Pre-Processing Tasks in Indonesian Twitter Messages. *Journal of Physics: Conference Series, 801*(1), 1-6.

Hull, D. A. and Grefenstette, G. (1996). A Detailed Analysis of English Stemming Algorithms. *Rank Xerox research Centre, 6*, 1-16.

Idris, N. and Mustapha, S. S. (2001). Stemming for Term Conflation in Malay Texts. *International Conference of Artificial Intelligence (ICAI 2001)*, Las Vegas, USA, 1512–1517.

Jabbar, A., Iqbal, S., Khan, M. U. G. and Hussain, S. (2018). A Survey on Urdu and Urdu Like Language Stemmers and Stemming Techniques. *Artificial Intelligence Review, 49*(3), 339-373.

Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AlZu'bi, S., Elbes, M., Hawashin, B. and Alia, M. A. (2019). A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan: IEEE, 622-628.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K. and Nerur, S. (2018). Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers, 20*(3), 531-558.

Kasthuri, M. and Kumar, D. B. R. (2014). A Comprehensive Analyze of Stemming Algorithms for Indian and Non-Indian Languages. *International Journal of Computer Engineering & Applications (IJCEA), 7*(3), 1-8.

Kaur, P. and Buttar, P. K. (2018). Review on Stemming Techniques. *International Journal of Advanced Research in Computer Science, 9*(5), 64-68.

Khan, R. U., Mohamad, F. S., UlHaq, M. I., Adruce, S. A. Z., Anding, P. N., Khan, S. N. and Al-Hababi, A. Y. S. (2017). Malay Language Stemmer. *International Journal for Research in Emerging Science and Technology, 4*(12), 1-9.

Kraaij, W. and Pohlmann, R. (1996). Viewing Stemming as Recall Enhancement. *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland: ACM New York, NY, USA, 40-48.

Kumar, N., Mitra, S., Bhattacharjee, M. and Mandal, L. (2019). Comparison of Different Classification Techniques Using Different Datasets. *Proceedings of International Ethical Hacking Conference 2018*, Kolkata, India: Springer Singapore, 261-272.

Lee, J., Othman, R. M. and Mohamad, N. Z. (2013). Syllable-Based Malay Word Stemmer. *2013 IEEE Symposium on Computers and Informatics (ISCI)*, Langkawi, Malaysia: IEEE, 7-11.

Leong, L. C., Basri, S. and Alfred, R. (2012). Enhancing Malay Stemming Algorithm with Background Knowledge. *Pacific Rim International Conference on Artificial Intelligence*, Kuching, Malaysia: Springer Berlin Heidelberg, 753-758.

Liliana, D. Y., Hardianto, A. and Ridok, M. (2011). Indonesian News Classification Using Support Vector Machine. *World Academy of Science, Engineering and Technology, 57*, 767-770.

Lovins, J. B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics, 11*(1-2), 22-31.

Magriyanti, A. A. (2018). Analisis Pengembangan Algoritma Porter Stemming Dalam Bahasa Indonesia. *Majalah Ilmiah Teknologi Elektro, 18*(2). doi:10.31227/osf.io/7ge4v

Marzuki, E. (2013). Linguistic Features in Sms Apologies by Malay Native Speakers. *GEMA Online® Journal of Language Studies, 13*(3), 179-192.

Muhamad, N. A., Idris, N. and Saloot, M. A. (2017). Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization. *Journal of Physics: Conference Series, 806*(1), 1-7.

Musa, H., Kadir, R. A., Azman, A. and Abdullah, M. T. (2011). Syllabification Algorithm Based on Syllable Rules Matching for Malay Language. *Proceedings of The 10th WSEAS International Conference on Applied*

Computer and Applied Computational Science, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 279-286.

Othman, A. (1993). *Pengakar Perkataan Melayu Untuk Sistem Capaian Dokumen.* (Master Thesis), Universiti Kebangsaan Malaysia, Bangi, Selangor,

Paice, C. D. (1994). An Evaluation Method for Stemming Algorithms. *SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland: Springer-Verlag London, 42-50.

Pennell, D. L. and Liu, Y. (2014). Normalization of Informal Text. *Computer Speech and Language, 28*(1), 256-277.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program, 14*(3), 130-137.

Pramono, L. H. and Subiantoro, C. (2018). Pengaruh Stemming Terhadap Ekstraksi Topik Menggunakan Metode Tf* Idf* Df Pada Aplikasi Pds. *JIKO Jurnal Informatika dan Komputer, 2*(1), 15-23.

Rais, N. H., Abdullah, M. T. and Kadir, R. A. (2010). Query Translation Architecture for Malay-English Cross-Language Information Retrieval System. *2010 International Symposium on Information Technology, 2*, 990-993.

Rajput, B. S. and NilayKhare, A. (2015). A Survey of Stemming Algorithms for Information Retrieval. *IOSR Journal of Computer Engineering, 17*(3), 76-80.

Ranaivo-Malancon, B. (2004). Computational Analysis of Affixed Words in Malay Language. *Proceedings of the 8th International Symposium on Malay/Indonesian Linguistics*, Penang, Malaysia, 50-59.

Rani, S. R., Ramesh, B., Anusha, M. and Sathiaseelan, J. (2015). Evaluation of Stemming Techniques for Text Classification. *International Journal of Computer Science and Mobile Computing, 4*(3), 165-171.

Rizki, A. S., Tjahyanto, A. and Trialih, R. (2019). Comparison of Stemming Algorithms on Indonesian Text Processing. *Journal of Telecommunication Computing Electronics and Control, 17*(1), 95-102.

Saleh, A. L. and Fadzli, S. A. (2018). A Proposed Method for Reducing the Dimension of Arabic Documents. *International Journal of Engineering and Technology, 7*(3.28), 205-208.

Saloot, M. A. (2018). *Corpus-Driven Malay Language Tweet Normalization*: (Doctoral Dissertation), University of Malaya.

Saloot, M. A., Idris, N. and Aw, A. (2014a). Noisy Text Normalization Using an Enhanced Language Model. *Proceedings of the International Conference on*

Artificial Intelligence and Pattern Recognition, Kuala Lumpur, Malaysia: SDIWC, 111-122.

Saloot, M. A., Idris, N. and Mahmud, R. (2014b). An Architecture for Malay Tweet Normalization. *Information Processing and Management, 50*(5), 621-633.

Saloot, M. A., Idris, N., Shuib, L., Raj, R. G. and Aw, A. (2015). Toward Tweets Normalization Using Maximum Entropy. *Proceedings of the Workshop on Noisy User-Generated Text*, Beijing, China: Association for Computational Linguistics, 19-27.

Samsudin, N., Puteh, M. and Hamdan, A. R. (2011). Bess or Xbest: Mining the Malaysian Online Reviews. *2011 3rd Conference on Data Mining and Optimization (DMO)*, Putrajaya, Malaysia: IEEE, 38-43.

Samsudin, N., Puteh, M., Hamdan, A. R., Nazri, A. and Zakree, M. (2012). Normalization of Common Noisy Terms in Malaysian Online Media. *Proceedings of the Knowledge Management International Conference*, 515-520.

Samsudin, N., Puteh, M., Hamdan, A. R. and Nazri, M. Z. A. (2013). Normalization of Noisy Texts in Malaysian Online Reviews. *Journal of ICT, 12*, 147-159.

Sankupellay, M. and Valliappan, S. (2006). Malay Language Stemmer. *Sunway Academic Journal, 3*, 147-153.

Schofield, A. and Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics, 4*, 287-300.

Sembok, T. M., Bakar, Z. A. and Ahmad, F. (2011). Experiments in Malay Information Retrieval. *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia: IEEE, 1-5.

Silvello, G., Bucco, R., Busato, G., Fornari, G., Langeli, A., Purpura, A., Rocco, G., Tezza, A. and Agosti, M. (2018). Statistical Stemmers: A Reproducibility Study. *European Conference on Information Retrieval*, Grenoble, France: Springer Cham, 385-397.

Singh, J. and Gupta, V. (2017a). An Efficient Corpus-Based Stemmer. *Cognitive Computation, 9*(5), 671-688.

Singh, J. and Gupta, V. (2017b). A Systematic Review of Text Stemming Techniques. *Artificial Intelligence Review, 48*(2), 157-217.

Singh, J. and Gupta, V. (2019). A Novel Unsupervised Corpus-Based Stemming Technique Using Lexicon and Corpus Statistics. *Knowledge-Based Systems, 180*, 147-162.

Srividhya, V. and Anitha, R. (2010). Evaluating Preprocessing Techniques in Text Categorization. *International Journal of Computer Science and Applications, 47*(11), 49-51.

Subagyo, P. A. (2007). Ciri-Ciri Kreatif Bahasa Sms. *SINTESIS, 5*(2), 167-186.

Sugumar, R. (2018). Improved Performance of Stemming Using Efficient Stemmer Algorithm for Information Retrieval. *Journal of Global Research in Computer Science, 9*(5), 1-5.

Suhartono, D. (2014). Lemmatization Technique in Bahasa Indonesian. *Journal of Software, 9*(5), 1202-1209.

Swain, K. and Nayak, A. K. (2018). A Review on Rule-Based and Hybrid Stemming Techniques. *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, Hunan, China: IEEE, 25-29.

Tai, S. Y., Ong, C. S. and Abdullah, N. A. (2000). On Designing an Automated Malaysian Stemmer for the Malay Language. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China: ACM Digital Library, 201–208.

Toman, M., Tesar, R. and Jezek, K. (2006). Influence of Word Normalization on Text Classification. *Proceedings of InSciT*, Merida, Spain: Open Institute of Knowledge, 354-358.

Willett, P. (2006). The Porter Stemming Algorithm: Then and Now. *Program, 40*(3), 219-223.

Yamout, F., Demachkieh, R., Hamdan, G. and Sabra, R. (2004). Further Enhancement to the Porter's Stemming Algorithm. *Machine Learning and Interaction for Text based Information Retrieval, Germany*, 7-23.

Yasukawa, M., Lim, H. T. and Yokoo, H. (2009). Stemming Malay Text and Its Application in Automatic Text Categorization. *IJCSNS International Journal of Computer Science and Network Security, 92*(12), 2351-2359.

Yeo, D. and Ting, S. H. (2017). Netspeak Features in Facebook Communication of Malaysian University Students. *Journal of Advanced Research in Social and Behavioural Sciences, 6*, 81-90.

Yunus, M., Zainuddin, R. and Abdullah, N. (2010). Semantic Query with Stemmer for Quran Documents Results. *2010 IEEE Conference on Open Systems (ICOS 2010)*, Kuala Lumpur, Malaysia: IEEE, 40-44.

Zeroual, I. and Lakhouaja, A. (2017). Arabic Information Retrieval: Stemming or Lemmatization. *2017 Intelligent Systems and Computer Vision (ISCV)*, 1-6.