Malay Language Stemmer

Rehman Ullah Khan*[1], Fitri Suraya Mohamad[2], Muh. Inam UlHaq[3], Shahren Ahmad Zadi Adruce[4], Philip Nuli Anding[5], Sajjad Nawaz Khan[6], Abdulrazak Yahya Saleh Al-Hababi[7]

1,2,4,5,6 Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

E-mail: krullah@unimas.my

E-mail: mfitri@unimas.my

E-mail: azshahren@unimas.my

E-mail: aphilip@unimas.my

E-mail: sajjadnawazkhan@gmail.com

E-mail: ysahabdulrazak@unimas.my

3Department of Computer Sciences, Khushal Khan Khattak University, Karak Pakistan

E-mail: inamix@gmail.com

ABSTRACT

Stemmer is a language processing tool that has been widely used in many artificial intelligence applications for removing affixes in a word such as prefixes, infixes, and suffixes to generate the root word. This study designs an algorithm and develops a Malay language stemmer. It is given that most of Malay language stemmers have problems in stemming, as they tended to have dependencies on online dictionaries, which return false results during stemming. It is given that the complexity of affixes in Malay words is higher than that of English words. Therefore, an offline dictionary of 9,512 words is introduced in this study to handle the ambiguity when stemming Malay words. Each step the algorithm first checks the word in the local dictionary as a root word, otherwise process the word. The five steps are stem-extra-suffix, stem-plural, stem-infix, stem-prefix, and stem-suffix. The affixes rules are extracted from Kamus Tatabahasa, and Kamus Dewan (4th Ed) is used to confirm the accuracy of stemmed words. The results show that the proposed stemmer can stem prefixes, suffixes and infixes with high accuracy. The study conclusively illustrated that the proposed stemmer can handle the complexity of Malay words. This stemmer can be further enhanced by a look-up table or dictionary of overlapping words to cover the prefix and suffix overlapping limitation.

Index Term — Stemming, Stemmer, Natural language processing, Algorithm and Morphology.

1. INTRODUCTION

The Malay language is well known part of Austronesian family which is spoken across South East Asia such as Malaysia, Indonesia, Singapore, and Brunei [1]. Bahasa Malaysia in Malaysia, Bahasa Indonesia in Indonesia and Bahasa Melayu in Singapore and Brunei, they originated from Malay language [1]. Malay language has two key features to take note of. First, it does not have grammatical functions as those commonly used in the English language. Second, affixes in Malay language analyses part-of-speech of the words to represent nouns, verbs and adjectives while affixes in English language represent plural, tenses and possession. The lack of morphological analyses published on the Malay language creates a need to address the gap in literature about the linguistic characteristics of the language widely used in Southeast Asia.

Stemming algorithm has been widely used today to serve various purposes. It is known as the process of eliminating derivational and inflectional suffixes from words until the root word is obtained [2]. For example, the words "assign", "assigns", "assigned", "assigning", "assignation" or "assignment" are reduced to the root word, "assign". Experiments were conducted to examine how efficient stemming is, and n-grams in identifying suffixes, multi-word concepts and spelling errors. The experiment was divided into bigram and trigram string matching using a document in Malay language. Thenceforth, Sembok and Zainab [3] [3] carried out separated experiments by using bigram and stemmed bigram as well as trigram and stemmed trigram. Their experiment revealed that stemming both keywords and documents has obvious advantage over stemming keywords only, and not stemming of any keywords. On the other hand, the experiment also revealed that bigram and trigram search worked better with no stemming on keywords. This is because when the keywords in the text are stemmed, the bigram and trigram search would be affected as well caused by the reduced keywords [3]. However, the experiment eventually showed that applying stemming to keywords and documents has improved the average precision value. Conclusively, Sembok and Zainab [3] have proven that retrieval effectiveness is improved by using combined search, n-gram matching and stemming.

Studies in the stemming algorithm for Malay language are relatively left behind in comparison to other languages such as English and European language [4]. The availability of Malay information retrieval system is also very limited. The usage of affixes in English and other European language is less complex than Malay language as it has been found that the stemmers are only concerned with the removal of suffixes. However, in Malay morphology, a stemmed word is produced by removing affixes in the text document or query. Affix is the verbal element that attached to the word whether at the beginning of the word (prefix) and at the end of the word (suffix). Besides, more than one affix may also be attached to a word at the same time. The word also can contain both affixes and this is known as prefix-suffix pair, for example as seen in the word 'pemakanan'. The root word for this word is 'makan' and the prefix is 'pe' is added at the beginning of the word and the suffix 'an' at the end of the word to complete the word 'pemakanan'. English and Malay languages differ in terms of their root words, which are based on their respective morphological structures [5]. For instance, the English words 'related', 'relates', and 'relation', are derived from the root word 'relate', and stemmer can work as suffix removal for English language. Yet, the Malay language has a different stemming process compared to English, due the complexity of its morphological rules. For example, the Malay words 'pengajaran', 'pembelajaran', and 'pelajar' are derived from the root word 'ajar', and it is insufficient to use suffix removal to decide for the perfect root word [6].

A Malay language stemmer that is used in text categorization was developed by Yasukawa et. al. [7]. This stemmer would check an input word with the dictionary before removing the affix to overcome the over stemming problem. In its methodology, the affixes are arranged from the longest match list to the shortest match list. In the longest match list, stemmer will remove the affix in the shortest match if there is no more root word after the affix is removed in the longest match. Therefore, the algorithm of this stemmer would not return a root word. However, it leads to two limitations in the stemmer. The limitations of the stemmer are ambiguity problem, and the algorithm is found to be more suitable than the arrangement of the longest or shortest match list. Such phenomenon occurs because when the stemmed word is found similar to root word, there is no further checking for the next possible affix.

Based on Kassim and colleagues [6], the affixation words are derived from the combination of affixes, clitics, and particle. He added that affixes can be classified into prefixes, suffixes, and confixes and infixes. The most universal prefixes of Malay are di+, ke+, se+, ber+, men+, pen+, ter+, and per+. The prefixes normally attached at the beginning of the root words. The part-of-speech of the root words does not change if mixed with inflectional prefixes like di+, ke+, and se+ [5]. For example, "diambil" (taken) is a derived word from prefix di+ and the root word "ambil" (synonymous to "take" in English). Both words are considered verbs. In contradiction, the part-of-speech of the root words do change if mixed with derivational prefixes [6]. For example, the word "pelayan" is a noun which derived from the mixture of layan (serving), verb, and prefix

pe+. For suffixes, the universal are +an and +i. The suffixes usually attached at the ending of the root words. Apart from prefixes, suffixes also have inflectional and derivational. For inflectional suffix, example such as kuasai (powered) is a verb which derived from the root word kuasa (power), verb, and suffix +i. For derivational: there are suffix, minuman (drink), noun, from combination of a verb minum (drinking), and suffix +an. For confixes, there are two types; inflectional and derivational; these do not and do change part-of-speech respectively. For instance, "hendak" (want) → "dihendaki" (wanted), do not change the part-of speech, whereas "pakai" (use) → "pemakaian" (usage) changes the part-of-speech. Malay language also have infixes like +el+, +em+, and +er+ which attached at the middle of the root words. For example, "telunjuk" (fingers) from the root word "tunjuk" (point). Thus, there are still many available sequences of affixes to be attached to the base words9.

In 2012, a number of researchers proposed different methods of Malay language stemming. One of the proposals is termed as the UniSZA stemmer, and it proposed 7 simple rules which leads to reduction in dictionary dependencies and lower processing cost [8]. Fadzli et al., (2012) defined the rules namely; check dictionary, check length, double words, prefix, suffix, change spelling and suffix-i. They developed and enhanced Malay prefixes library based on RAO (Rules Application Order) stemmer proposed by Fatimah in 1995 in the prefix step. The suffix step was using similar approach as the prefix. Constructed rules are arranged in five different ways of Arrangement A, B, C, D and E. Fadzli et al. (2012) conducted an experiment on all five arrangements and the results showed that Arrangement C: Double Words → Check Dictionary → Check Length → Prefix → Suffix → Change Spelling → Suffix-i scored well in terms of accuracy. In comparisons to RAO and RFO (Rules Frequency Order), UniSZA perform better with compression rate of 67.26%. On the other hand, they found that UniSZA method also improves other languages' stemmer compression rate.

An innovative approach of stemming called Malay stemmer with background knowledge is proposed by Leong et al., (2012). It was designed to avoid excessively broad-spectrum dictionary scanning of traditional algorithm, where words are scanned regardless if they have affixes, to boost the processing speed. A dictionary with affixed root words is additionally added as a reference before stemming to solve stemming error mentioned by Leong et al., (2012). Leong et al., (2012) proposed that this stemmer uses RFO as the basis algorithm. The only difference is the implementation of second dictionary which checks on affixed root words. The first word will be going through first basic dictionary of Kamus Dewan (4th Ed) to get the root word; if exists, it proceeds to the next word; if it does not exist, the second dictionary will be accessed Leong et al., (2012). When the second dictionary is accessed, there are three rules, (a) recode for prefix spelling exceptions and check the second dictionary, (b) check the stem for spelling variations and check the second dictionary and (c) recode for suffix spelling exceptions and check the second dictionary. If step (c) failed to stem a root word, the process will be looped. This approach successfully decreases the stemming error of 0.21% to 0.09%.

In addition, a stemming method of exhaustive affix stripping and a Malay Word Register are used to solve over-stemming, under-stemming errors, and to address ambiguity problem of determining correct root word [9]. By considering all possible word morphologies, the over-stemming and under-stemming error remover helps in looking for possible affix to be removed in all order classes for example: [prefix + root], [root + suffix] or [prefix +root +suffix]. Additionally, the ambiguity reducer addresses the ambiguity problem of the derivative words by referring to Malay Word Register to determine the correct root word. Darwis et al., (2012) mentioned in the results of their test on a proposed method, as the Malay Word Register may not contain all possible derivative words to solve all ambiguity cases, it is still practically useful and does contribute to the 99.8% accuracy of their stemmer.

Among all those existing Malay stemmers, Lee et al., [10] developed a syllable-based Malay word stemmer. Unlike traditional stemming process, the proposed concept is to split the word into syllable set before stemming it. After the syllabification is done, stemming rules are used to identify the morphological structure of the words. In this research, there are three set of rules, namely Prefix rules, Suffix rules and Morphographemic rules. The prototype works by removing identified prefixes and suffixes, then consider spelling variations and exceptions. However, limitations still exist in

the research as different words are under-stemmed or over-stemmed when accorded to the three rules. For instance, stemming result of peralatan was ralat while the root word is alat whereas stemming result of kediaman was dia while root word is diam [10]. These examples indicated the syllabification process might affect the accuracy of result of the syllable-based Malay word stemmer. Regardless of these weaknesses, the system recorded an achievement of a 97.4% of accuracy in stemming Malay words.

Singh and Gupta [11] described a comprehensive literature relevant to text stemming by classifying it according to certain key parameters; then it describes the deep analysis of some well-known stemming algorithms on standard data sets. Kassim et al., [6] presented a detailed review of Malay word stemmers. They explained the research trends of the existing Malay word stemmers based on morphological structures of Malay language, general word stemming methods and adopted word stemming. A cross-lingual sentiment lexicon acquisition method for the Malay and English languages is reported by Nasharuddin, et al., [12]. They further tested their algorithm on a set of news test collections. Knowles Gerry [11], proposed new standards of data collection, organisation and analysis associated with the methodology of corpus linguistics. A rule based algorithm by which a stem for the Arabian Gulf dialect can be defined [11]. Special rules are created to remove the suffixes and prefixes of the dialect words. Also, the algorithm applies rules related to the word size and the relation between adjacent letters.

In conclusion, there have been many approaches proposed and implemented by researchers in the past few years on Malay stemmers. Each reviewed approach has its own advantages and limitations, but the main contention lies in ensuring the Malay prefixes and suffixes are clearly represented into the system as they might affect the outcome. It is clear that, the stemming algorithm is still opened to various methods in modifying and improving the results of stemming. The contention of this paper is to provide another methodological perspective in designing a stemming algorithm and developing a stemmer for stemming Malay words, by using an offline dictionary to handle ambiguity during the stemming process.

2. MATERIALS AND METHODS

2.1. System Design

Stemming Malay language is not as easy as just by removing the suffix because Malay affixes consist of four diverse types which are:

Prefix – attached at the beginning of a word

Suffix – attached at the end of a word

Infix – located at the middle of a word

Prefix-suffix pair (confix) – attached at the beginning and the end of the word.

It has already been established that the Malay language is more complex than the English language. For the purpose of this study, an offline dictionary of 9,512 words is created to handle ambiguity during the stemming process. The affixes rules included within the algorithm are extracted from Kamus Tatabahasa and the dictionary used to check the validity of a stemmed word, is Kamus Dewan (4th Ed). The system used to produce the stemmer is HP core i7 and Windows 10 operating system. The tools used are Python language, Anaconda, NLTK and Pycharm (community version) to develop the stemmer. Fig.1 shows the dataflow of the proposed algorithm.

2.2. Algorithm

The stemmer is based on the following algorithm. This algorithm has five steps. Each step checks the input word or stemmed word in local dictionary. If the word is root word, then the word is printed as a root word otherwise the word is processed to stem according to the defined rules. The pseudo code of the algorithm in given in Fig. 2 below.

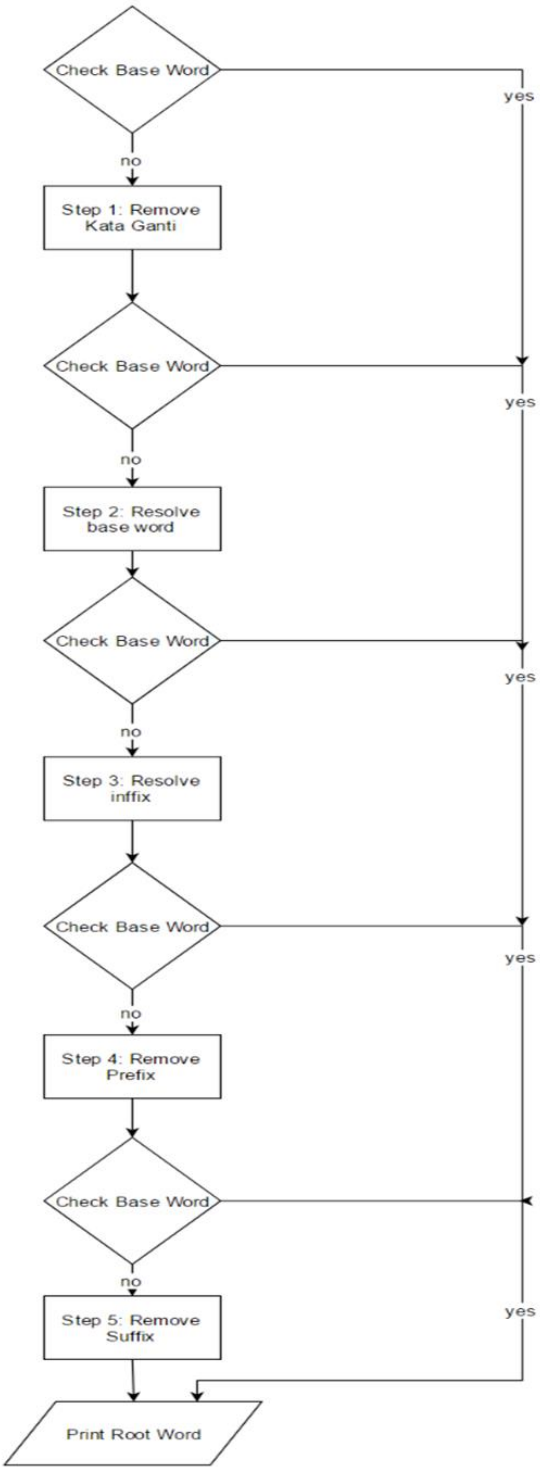

Fig-1: The dataflow of our stemming process

While not stop equal to yes do:
Get input word or paragraph
Check in Dictionary
Check input/stemmed word in offline dictionary as a root word
if the word is root word
Print the word as root word.
Do you want to stop, Yes/No?
If stop equal to yes
STOP
Endif
else Go to next Step
endif

Step 1 Stem_extraSuffix

If the word contains extra suffix which is –nya,
then remove it
Check stemmed word in dictionary.
endif

Step 2 Stem_Plural

If the word is in plural form,
then remove plural.
Check stemmed word in dictionary.
endif

Step 3: (Stem_infix

If the word contains infix,
then remove infix.
Check stemmed word in dictionary.
endif

Step 4: (Stem_prefix)

If the word contains prefix,
then remove prefix.
Check stemmed word in dictionary.
endif

Step 5: (Stem_suffix)

If the word contains sufix,
then remove sufix.
Check stemmed word in dictionary.
endif
end While

Fig- 2: The pseudo code of the algorithm.

First, the input word or stemmed word in each step is checked in the local dictionary. If the word is found as root word, then the word will be displayed as root word. Otherwise, the process will proceed to next step.

Step1: (Stem_extraSuffix),

It would stem the extra suffix which is "–nya". Without using the stem "–nya" at the first step, the root word is a meaningless word. For example, for the word 'mendekatinya'.

Execute the step without step1: (stem_extraSuffix)

After (check root_word) and (stem_infix), then

Stem prefix: "dekatinya"

Stem suffix: “dekati”

Execute step with step1: (stem_extraSuffix)

Stem extraSuffix: “mendekati” (“-nya” would be stem first)

Then proceed to (check root_word) and (stem_infix), after that

Stem prefix: “dekati”

Stem suffix: “dekat”

Accordingly, it is necessary to include stem_extraSuffix at the beginning in order to prevent any meaningless root word.

Step2: (Stem_Plural),

Malay language has different mechanism for making plurals. The particular word is doubled to make plural, for example buku-buku (books). In this step the word is examined. If the word is plural, then it is stemmed to root word.

Step3: (Stem_infix),

If the word contain infix, then infix is removed this step and proceed to next step.

Step4: (Stem_prefix),

The prefixes are removed in this step for example “diper”, “ber”, “per”, “ter”, and so forth. However, there is a special grammar present in the Malay language where a word would be replaced with a different letter for different prefixes. Therefore, prefix “mem” would be replaced with either the letter “f” or “p” after checking with the dictionary. For example, “memakai” would become “pakai” and “memikir” would become “fikir” after stemming. The step would check the word if there are less than four alphabets exist before stemming, to prevent the loss of meaning for the word. It would not remove the prefix if the stem is too short.

Step5: (stem_suffix),

The suffixes are removed in this step for example “kannya”, “nya”, “kan”, “an”, etc. Like step4, it would not remove the suffix if the stem is too short. It also checks the word whether less than five alphabets before stemming to prevent loss of meaning for the word.

3. RESULTS AND DISCUSSION

The stemmer can remove prefixes, suffixes and infixes from words in order to obtain the root word. To test the performance of the stemmer a Malay language essay is randomly chosen from an online source [13]. A sample result is shown in Fig. 3 below.

```
>>> p.stem("patriotik")
'patriotik'
>>> p.stem("semakin")
'makin'
>>> p.stem("luntur")
'luntur'
>>> p.stem("dalam")
'dalam'
>>> p.stem("kalangan")
'kalang'
>>> p.stem("masyarakat")
'masyarakat'
>>> p.stem("di")
'di'
>>> p.stem("negara")
'negara'
>>> p.stem("kita")
'kita'
>>> p.stem("pada")
'pada'
>>> p.stem("masa")
'masa'
>>> p.stem("ini")
'ini'
>>> p.stem("seperti")
'seperti'
>>> p.stem("yang")
'yang'
>>> p.stem("didakwa")
'dakwa'
>>> p.stem("oleh")
'oleh'
>>> p.stem("banyak")
'banyak'
>>> p.stem("pihak")
```

Fig- 3: Sample test of stemming

The following Table 1 shows a sample list of prefixes, suffixes that can be stemmed using this stemmer.

Table-1: Sample list of removable prefixes and suffixes

Prefixes	"diper", "ber", "bel", "per", "ter", "mem", "penye", "peny", "menye", "meny", "menge", "penge", "meng", "peng", "men", "pen", "me", "pem", "pe", "be", "ke", "se", "ter", "te", "di"
Suffixes	"kannya", "nya", "kan", "an", "i", "kah", "lah", "pun", "ita", "man", "wan", "wati", "ku", "mu"

Other than prefixes and suffixes, the proposed stemmer can stem infixes also. In the Malay language, there is a unique rule called dual words or “kata ganda”. There are several instances of “kata ganda” that exist in the Malay language, the proposed stemmer can successfully stem “kata ganda”. The following Table 2 shows several examples of dual word stemming.

Table-2: Sample list of dual words stemming

Type of Dual Words	Words	Stemmed Word
Words without prefix	jalan-jalan	jalan
Non-identical words without prefix	saudara-mara	saudara
Identical words with prefix	tergesa-gesa	gesa
Non-identical words with prefix	membeli-belah	beli
Words with suffix	barang-barangan	barang
Words with prefix and suffix	sebaik-baiknya	baik

The stemmer is also able to stem passages instead of just words. Words make up sentences and sentences make up a passage. Therefore, any text passage could be stemmed using the stemmer and output as text passages but with stemmed words. Fig. 4 and Fig. 5 below show the output for stemmed text passages with and without the use of local dictionary.

- Text passage without dictionary

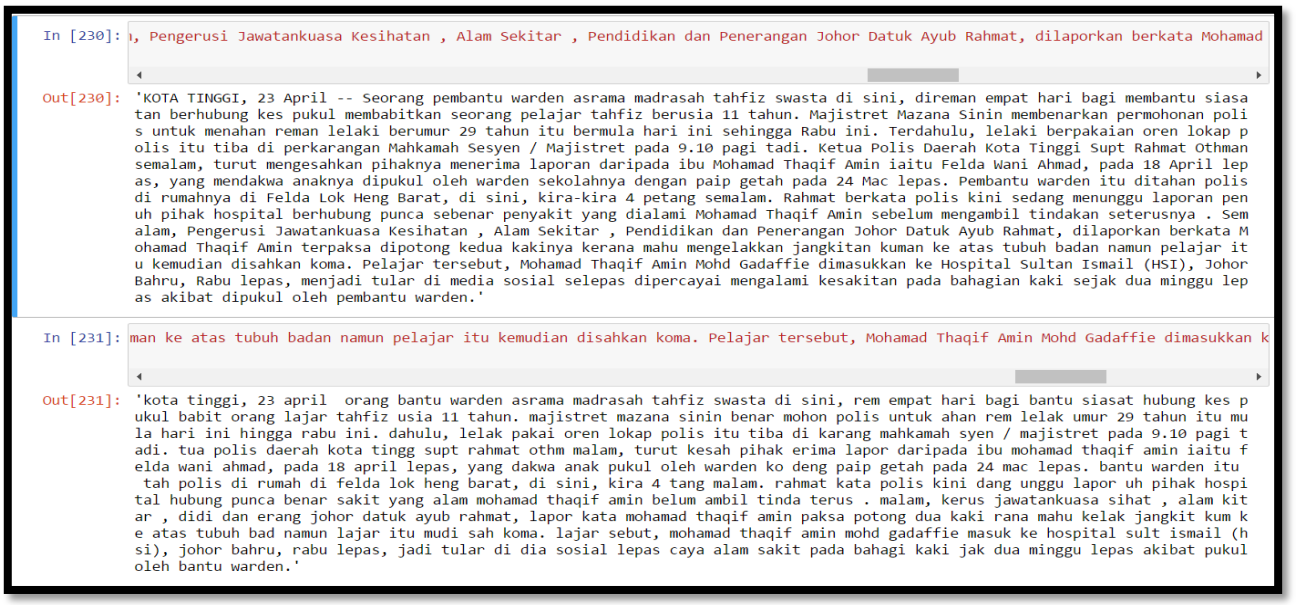

Fig- 4: Output for a stemmed text passage without the use of word dictionary

- Text passage with dictionary

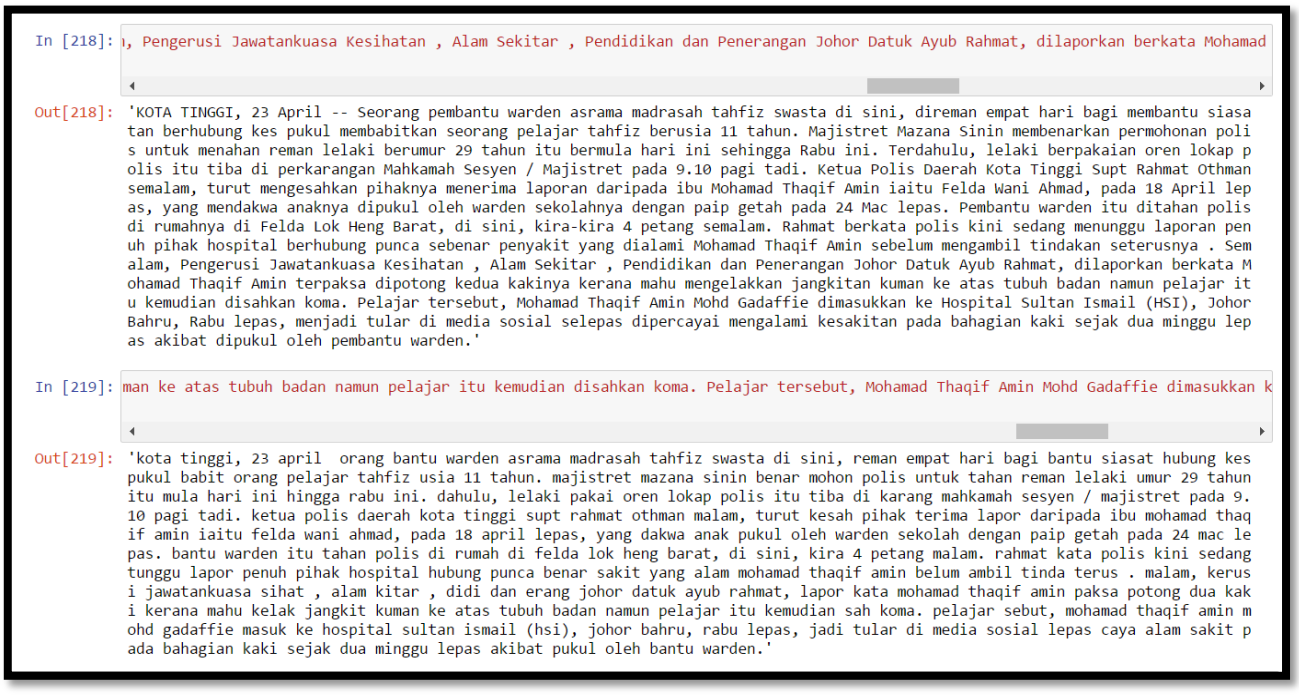

Fig- 5: Output for a stemmed text passage with use of word dictionary.

The accuracy of stemming increased tremendously with the use of word dictionary. In Figure 4, it is shown that without the help of a words dictionary, a handful of words were incorrectly stemmed as they contained letters that resembled prefixes or suffixes. However, they are actually a part of the root word, therefore the stemmer incorrectly stems the words and turns them into meaningless words. But Figure 5 shows that the same words can be accurately stemmed with the help of local dictionary. It can be concluded that the use of a word dictionary is essential in improving the accuracy of the stemmer provided the dictionary contains many different words for the stemmer's reference.

The stemmer has its limitations. It does not achieve a hundred percent accuracy. Similar to the Porter stemming algorithm, there are instances where several words were not properly stemmed as the root words contain letters which were also found in the prefixes, therefore causing an overlap. The stemmer always goes for the longer prefix or suffix such as "pem" instead of "pe" which is the longer prefix among the two. As an example, the word "pemain" contains the prefix "pe" while its rootword consists of "main". However, the stemmer recognizes the prefix "pem" instead of "pe" and this results in the word being improperly stemmed with an output of "ain" as a result.

The same problem arises when an ending letter of a root word overlaps with letters found on suffixes as well. An example of this is the word "pendidikan" where the root word which is "didik" has a letter "k" as an ending letter and it overlaps with the suffix "kan" where it is supposed stem the suffix "an" only. Hence, the output of the word becomes "didi" as the stemmer always stem the longer suffix from words.

4. CONCLUSION

It is clear in the study that the proposed stemmer is able to accurately stem extra suffix, Malay plural, prefix, infix and suffix. It is recommended that this stemmer can be further enhanced by look up table or dictionary of overlapping words to cover the prefix and suffix overlapping limitation.

REFERENCES

[1] Abdullah, M. T., Ahmad, F., Mahmod, R., & Sembok, T. M. T. [3]. Rules frequency order stemmer for Malay language. IJCSNS International Journal of Computer Science and Network Security, 9[14], 433-438.

[2] Anelyza. [13]. Efforts to enhance the patriotic spirit among the communities in our country. Retrieved from, Ranaivo-Malancon, B. *Computational analysis of affixed words in malay language*. in *Proceedings of the 8th International Symposium on Malay/Indonesian Linguistics (ISMIL)*. 2004. Penang, Malaysia.

[3] Lovins, J.B., *Development of a stemming algorithm.* Mech. Translat. & Comp. Linguistics, 1968. **11**(1-2): p. 22-31.

[4] Sembok, T.M.T. and Z.A. Bakar, *Effectiveness of stemming and n-grams string similarity matching on Malay documents.* International Journal of Applied Mathematics and Informatics, 2011. **5**(3): p. 208-215.

[5] Abdullah, M.T., et al., *Rules frequency order stemmer for Malay language.* IJCSNS International Journal of Computer Science and Network Security, 2009. **9**(2): p. 433-438.

[6] Kassim, M.N., et al. *Word stemming challenges in Malay texts: A literature review*. in *4th International Conference on Information and Communication Technology (ICoICT)*. 2016. Bandung, Indonesia: IEEE.

[7] Kassim, M.N., et al. *Malay Word Stemmer to Stem Standard and Slang Word Patterns on Social Media*. in *Tan Y., Shi Y. (eds) Data Mining and Big Data. DMBD 2016. Lecture Notes in Computer Science*. 2016. Cham Springer.

[8] Yasukawa, M., H.T. Lim, and H. Yokoo, *Stemming malay text and its application in automatic text categorization.* IEICE transactions on information and systems, 2009. **92**(12): p. 2351-2359.

[9] Fadzli, S.A., et al. *Simple rules malay stemmer*. in *The International Conference on Informatics and Applications (ICIA2012)*. 2012. Malaysia: The Society of Digital Information and Wireless Communication.

[10] Darwis, S.A., R. Abdullah, and N. Idris, *Exhaustive affix stripping and a Malay word register to solve stemming errors and ambiguity problem in Malay stemmers.* Malaysian Journal of Computer Science, 2012. **25**(4): p. 196-209.

[11] Lee , C.J., M.O. Rosita, and N.Z. Mohamad. *Syllable-based Malay Word Stemmer*. in *2013 IEEE Symposium on*

Computers & Informatics (ISCI) 2014. Langkawi, Malaysia: IEEE.

[12] Knowles Gerry, *Languages and linguistics in 2003: The potential contribution of corpus linguistics.* Journal of Modern Languages, 2017. **V. 15**(1): p. 37-50.

[13] Nasharuddin, N.A., et al. *English and Malay Cross-lingual Sentiment Lexicon Acquisition and Analysis.* in *Kim K., Joukov N. (eds) Information Science and Applications 2017. ICISA 2017. Lecture Notes in Electrical Engineering, vol 424.* 2017. Singapore: Springer.

[14] Anelyza, *Efforts to enhance the patriotic spirit among the communities in our country*, in *Teacher Anelyza SPM* 2009.

[15] Rehman Ullah Khan1*, M.I., YahyaKhan3,Oon Yin Bee4, Shahren Ahmad ZadiAdruce5, Mai S. Ishak6, Tan KockWah7, *A NOVEL ALGORITHM FOR TEXT STEGANOGRAPHY.* International Journal of Soft Computing: p. 13.

