INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN: 2289-8522 e-ISSN: 2180-0650
VOL. 9, ISSUE 1, 39 – 45
DOI: https://doi.org/10.15282/ijsecs.9.1.2023.4.0108

ORIGINAL ARTICLE

Extraction of Malay Root Word that Starts with Letter P in Malay e-Khutbah using Rule Based

Nurhilyana Anuar [1], Zamri Abu Bakar [1,*], Normaly Kamal Ismail [2]

[1]Centre of Foundation Studies, Universiti Teknologi MARA Cawangan Selangor Kampus Dengkil, 43800 Selangor, Malaysia.
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Selangor, Malaysia.

ABSTRACT – Stemming is an important process in text processing especially in Natural Language Processing (NLP). It could extract root word from the affix words in the text. In addition, it helps in extracting useful information that contributes to many area of research study such as Information Retrieval. Several stemming algorithms have been discussed in previous studies. However, there are limited studies on Malay stemming process and the number of experimental data used. In this study, we focus on stemming process of Malay stemming algorithm by using rule-based algorithm for a larger dataset of Malay language text. The syntatic linguistic rule-based method was used in the stemming process involves of removing prefixes, suffixes and, prefixes and suffixes. Training dataset was used in this study which consisted of 3233 sentences from e-khutbah text. The result of the experimental evaluation was done by measuring the precision, recall and f-measure. It was found that the algorithm used in this study showed a promising result based on total of dataset used for each test. The value of precision, recall and F-measure icrease to 95%, 97% and 97% respectively. The enhancement of the stemming process has shown a significant impact on Malay text processing which in general improved the performance of NLP applications..

ARTICLE HISTORY
Received: 27 October 2022
Revised: 21 December 2022
Accepted: 18 February 2023
Published: 25 February 2023

KEYWORDS
Stemming
Affix
Natural language
Processing
Rule-based approach
Root word

INTRODUCTION

Stemming process is an important stage in Natural Language Processing (NLP), where it will contribute to the knowledge extraction. Furthermore, stemming is useful in text processing especially to extract information in the text document. Stemming stage is commonly used in text-based applications such as search engines, machine language translators, chatbots and spelling checkers [1]. The process of transforming the derivative word into root word is called stemming. The process of stemming involves eliminating prefixes, infixes, and suffixes of derivation words until it obtains the root word. For example, English words like “bake”, “bakes”, “baking” or “baked” are derived from the root word “bake”. Different languages have different rules of stemming the word because each of the languages has its own morphology. However, the Malay word has complex morphological rules [2]. For example, the Malay word “membakar” (burning), “pembakaran” (burning), “pembakar” (burner), “kebakaran” (fire) or “terbakar” (on fire) are derived from the root word “bakar” (burn).

The first study on Malay stemmer algorithm was conducted by Othman who developed a rule-based approach which came out with 121 morphological rules. However, some of the rules were not stemmed correctly and produced ambiguous meaning of root word [2]. Since then, many studies have been conducted to improve the stemming algorithm in Malay text document. Several challenges have been highlighted such as morphological rules of Malay language, word patterns in digital Malay texts and word stemming evaluation in producing effective word stemming algorithm of Malay language [2]. Therefore, this study is conducted to develop a model using the proposed algorithm in order to improve the process of stemming. Even though there are many discussions on the stemming process, there are still lack of studies on stemming for Malay language. In most studies, there is a limited number of data being experimented. In order to obtain a better level of accuracy for the stemming algorithm, this study proposed to improve the algorithm by using rule-based approach for the Malay text document. As such, this study used e-khutbah as the data for this experiment and the stemming process applied the words which started with letter ‘p’. The data were selected due to the reason that there are still lack of studies to test this type of data. In addition, the letter ‘p’ was selected because its frequency is higher than other letters in the dataset. Furthermore, this study also evaluated the accuracy of the result by comparing it with previous studies. This study is important to summarize the document and it will help the people to understand the overall e-khutbah text.

RELATED WORK

Stemming is a process of extracting root word from derivation word in a text document. This is an important process in Information Retrieval (IR) field. Much research has been conducted in stemming algorithm which has been applied into several languages like English, Arabic, Indonesia, Kazakh, Malay, and few others. This algorithm has been

*CORRESPONDING AUTHOR | Zamri Abu Bakar | ✉ zamri@uitm.edu.my

implemented in several applications such as query word searching, plagiarism detection and others because it makes applications more efficient and effective [3][4]. Several stemming techniques have been discussed widely by researchers in this field such as rule-based, statistical, and hybrid approaches [5]. In early research of stemming algorithms, many papers discussed on English language morphology. Previous research used the rule-based approach as an algorithm for the stemming process. It is a process of stripping prefix and suffix of a word that is based on the morphology of the language to deduce the root word. The stemming algorithm for Kambaata language was based on rule-based approach which was similar to the Lovin's stemmer where the algorithm included single pass, context-sensitive, and longest-matching designed. The output from the stemmer indicated that 96.87% words were stemmed correctly, 2.60% words over stemmed and 0.54% words under stemmed. In addition, this stemmer was extended from Lovin's stemmer. After the stemming process was completed, it found that only 65.86% existed in the dictionary [6]. There are errors found in this stemmer due to the complexity of Kambaataa morphology.

Porter stemmer is an algorithm that uses a rule-based approach for stemming the words. Many languages have adapted this algorithm in previous studies such as Indonesia, English, Malay and others. Studies conducted by Arif Siswandi adapted Porter algorithm for Indonesian text documents. The type of stemming used was algorithm for dictionary-based. The experiment was carried out by using 100 predetermined Indonesian text documents. The result revealed the highest accuracy of Porter algorithm in which 94% were stemmed correctly, 4 % were over stemmed and 0.9 % were under stemmed [7]. The weakness found when Porter algorithm was adapted in this study is that the word was not found in the dictionary since it was treated as root word. Stemming process is an important step before proceeding into further data analysis process. A study was conducted by [8] to make a comparison between two stemming algorithms which are Porter and Lancaster. The objective of the study was to find the stemming error produced by Porter and Lancaster algorithm. Based on the result, the Porter algorithm performed 43% better on stemming error produced as compared to Lancaster algorithm [8]. Even though Porter algorithm performed better than Lancaster in this paper, stemming error produced was still high. According to Maheswari, stemming is an important process of converting words in text to root word. The study proposed morphological variation removable stemming algorithm to improve the existing algorithms such as Porter and Lancaster. The experiment was conducted to compare the efficiency among Porter, Lancaster and the proposed morphological variation stemmer. The result revealed that Porter stemmer correctly stemmed by 71.01%, Lancaster stemmer was 47.71% and the proposed stemmer was 89.88%. Thus, the proposed stemmer had a higher correct stemmed word as compared to other existing algorithms [5]. Therefore, when the algorithm is able to produce higher correct stemmed word, stemming error was less compared to the existing algorithm of the study.

Malay language stemmer was proposed by [9] to handle the complexity of Malay word morphology. This stemmer uses rule-based approach and is able to remove the prefixes, suffixes, infixes and dual word ("kata ganda") to produce the root word. The result revealed that the stemmer can stem the prefix, suffix and infix word with high accuracy [9]. Even though this stemmer can stem with high accuracy, there is still limitation on overlapping words for prefixes and suffixes. There are many methods discussed by the scholars of NLP to stem the word by removing the prefixes, suffixes and infixes in variety of languages morphology in order to extract the root word. It is clear from the related work discussed above that the efficiency of rule-based approach produces high accuracy of stemming result. A few studies have implemented rules-based approach to the Malay language morphology. Thus, the result of stemming revealed a very compromising way to apply rule-based approach for Malay text document of this study.

METHODOLOGY

The process of stemming is discussed in further detail in this section. In order to correctly produce the expected results, several activities were performed in the research methodology section. Data went through pre-processing process such as tokenization, stop word removal and data cleaning. After the cleaning process was completed, the program filtered all the words starting from the letter 'p'. This is because the frequency of letter 'p' is higher than other letters in the dataset. Based on the sample data used, we found that there were about 500 words starting from letter 'p' which excluded redundant and stop words. The data were collected from online khutbah of Jabatan Agama Islam Selangor (JAIS) portal. The sample Malay text document used in this study had 3233 sentences taken from this online khutbah.

Stemming Processing

The Malay language word consists of 4 types of affixes which are prefix, infix, suffix, and confix. Prefix is the additional word located at the beginning of a root word. Suffix is the additional word located at the end of a root word. Infix is additional word at the middle of a root word. Lastly, confix is the additional word at the beginning and the end of a root word [5]. The development of rule-based algorithm used in this study was Python language.

Algorithm

Figure 1 shows the flow of the research methodology for this study. It consists of 6 steps to get the result which included the list of root words. Before entering the stemming process, it involves 4 steps including data cleaning process. It starts with importing Malay text document in the database, followed with the step of removing punctuation and digits before the tokenization process takes place. Next, the sentences are tokenized into a single word and followed by normalization process. In normalization process, the program eliminates all the stop words in the Malay language. The next step is filtering the data that starts with letter 'p' only. After finishing the previous step, the program starts with stemming process until it produces the result which is a list of root word. The words that have been tokenized are known as token. The process of stemming is then further discussed.

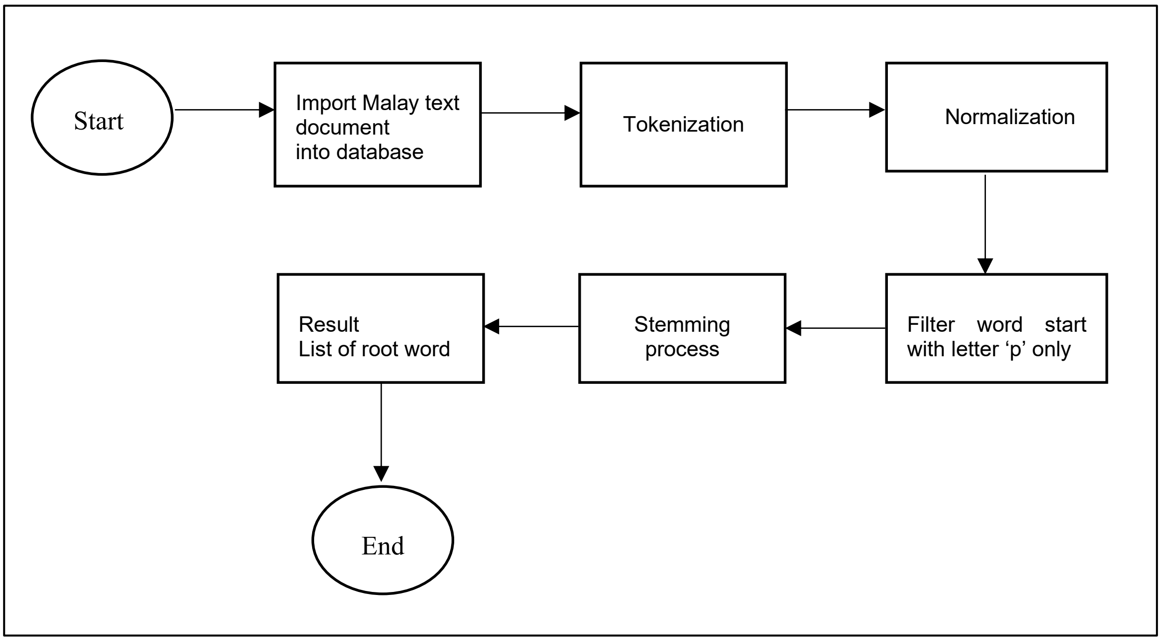
Start
Import Malay text document into database
Tokenization
Normalization
Filter word start with letter 'p' only
Stemming process
Result List of root word
End

Figure 1. Research methodology of study.

After filtering the words that start with letter 'p', the next process is to find the token whether it is as an affix or root word. In this process, the program checks the length of token that is less than 6 characters and compares it with the word in the dictionary. Based on the dataset, all the words that contain 6 characters are found as the root word and it is confirmed with the dictionary. If the token is found in the dictionary, therefore the status of the token is updated as pure root word. The algorithm is shown in Figure 2.

Algorithm 1: Pure root word from token

```
While token exist
    If token length less than 6 characters
      If token equal to word in dictionary
          Update status token to root word
End While
```

Figure 2. Pure root word from token.

Figure 3 shows the process of stemming for the prefix word from the existing token. It starts by checking the length of each token that is greater than 6 characters and it begins to find the word that starts with "pe", "pen", "per", "pem", "peng", "penge" and "peny". It then removes the prefix as in the rules based. The status of the token is updated in database as root word and the process continues until the end of file.

Algorithm 2 : Stem prefix word from token

```
While token exist
    If token length greater than 6 characters
        If token equal to "Pe" or "Pen" or "Per" or "Pem" or "Peng" or "Penge" or "Peny"
          Remove characters from token
          Store as root word in database
End While
```

Figure 3. Stem prefix word.

Figure 4 shows the process of stemming for the suffix word from the existing token. It starts by checking the length of each token which is greater than 6 characters and it begins to search words with "i", "an", "kan", "nya". It then removes the suffix as in the rules based. The status of the token is updated in the database as root word and the process continues until the end of the file.

```
Algorithm 3 : Stem suffix word from token

While token exist
    If token length greater than 6 characters
            If token  end with equal to "i" or "an" or "kan" or "nya"
            Update status token to affix word
      Else
            Update status  token to non-affix word
End While
```

Figure 4. Stem suffix word.

All the affixes tokens go through the process of stemming. Algorithm 4 in Figure 5 shows the process involved in stemming for this study. Firstly, the length of token is checked. If it is greater than 6 characters, the next process is to find the beginning characters of token that start with "pe", "pen", "pem", "per", "peng", "penge", "peny" and the token which ends with characters "i" or "an" or "kan" or "nya". If the token matches the rule used in Figure 3, the character is removed from the token and stored in the database as root word.

```
Algorithm 4: Stemming process

While token equal to affix word
  If token length greater than 6 characters
    If first two characters for token equal "Pe" AND last characters "i" or last two characters "an"
        or last three characters "kan" or last three characters "nya"
           Remove characters from token
           Store as root word in database

    If first three characters for token equal "Pen" or "Pem" or "Per" AND last characters "i" or
       last two characters "an" or last three characters "kan" or last three characters "nya"
           Remove characters from token
           Store as root word in database

    If first four characters for token equal "Peng" or "Peny" AND last characters "i" or last two
        characters "an" or last three characters "kan" or last three characters "nya"
             Remove characters from token
           Store as root word in database

    If first fifth characters for token equal "Penge" AND last characters "i" or last two characters
       "an" or last three characters "kan" or last three characters "nya"
          Remove characters from token
          Store as root word in database

   Next token
End While
```

Figure 5. Stem affix word.

Accuracy measure of the result

This study measures the accuracy of the result by assessing the performance of stemming algorithm through the calculation of precision, recall and f-measure. The equations used are as follows.

$$\boldsymbol{Precision} = \frac{\text{Correct Stemmed Cases}}{(\text{Correct Stemmed Cases} + \text{False Stemmed Cases})} \qquad (1)$$

$$\boldsymbol{Recall} = \frac{\text{Correct Stemmed Cases}}{(\text{Correct Stemmed Cases} + \text{Not Stemmed Cases})} \qquad (2)$$

$$\boldsymbol{F-Measure} = \frac{(\text{Precision x Recall })}{(\text{Precision} + \text{Recall})} - 1 \qquad (3)$$

Precision equation is used to compute corrected stemmed cases among the whole words which are correct stemmed cases and false stemmed cases of this experiment. Meanwhile, recall equation is used to assess the stemmed cases among the whole words used in this study. F-measure equation is applied to compute the mean precision and recall of stemmed cases to measure the performance.

RESULT AND ANALYSIS

Based on the rule-based approach used in this study, the result showed that it is able to remove the prefixes and suffixes starting with letter ‘p’. In addition, the algorithm used in this study is able to correctly stem and produce the root word in the Malay language of e-khutbah text. The sample result of words after stemming is shown in Tables 1, 2 and 3.

Table 1. Root words identified after stem prefix.

Stem Prefix: “pe”/ “pem” / “pen”/ “per”/ “peng”/ “penge”/ “peny”	
Original word (token)	**After stemming**
pekerja	kerja
pelekat	lekat
pelengkap	lengkap
penceroboh	ceroboh
pendakwah	dakwah
pembaca	baca
pembawa	bawa
penghasut	hasut
penghafaz	hafaz
penghapus	hapus

In Table 1, the words used in this study successfully stemmed the prefix words “pe”, “per”, “pem” , “peng” , “penge” and “peny” from the original word to root word such as “pekerja” (worker) to “kerja” (work), “pelekat” (sticker) to “lekat” (stick), “pembaca” (reader) to “baca” (read) and others.

Table 2. Root words identified after stem suffix.

Stem Suffix : “i”/ “an” / “kan”/ “nya”	
pedihnya	pedih
pendeknya	pendek
pentingnya	penting
perlunya	perlu
pegangan	pegang
pesanan	pesan
penjarakan	penjara

pentingkan	penting
perempuannya	perempuan

The result in Table 2 shows that the algorithm successfully stemmed the suffix ending with “i” , “an” , “kan” and “nya” and produced the correct root word. For example, the original word “perlunya” (necessity) becomes “perlu” (need), “pentingnya” (importantly) becomes “penting” (important) and others.

Table 3. Root words identified after stem prefix and suffix.

Stem Prefix and Suffix	
Original word (token)	**After stemming**
penilaian	nilai
penternakan	ternak
pembahagian	bahagi
pembuatnya	buat
pembentukkan	bentuk
perkembangan	kembang
permintaan	minta
perkatakan	kata
pertingkatkan	tingkat
penghujungnya	hujung

Next, Table 3 shows the experimental result of stemming prefix and suffix for each token based on the rule based. Some of the results successfully stemmed prefix and suffix which then produced the correct root words like “penilaian” (evaluation) becomes “nilai” (value), “pembahagian” (division) becomes “bahagi” (divide) and others. In this study, we collected about 3233 sentences from e-khutbah text. The result in Table 4 shows the performance of selected word used in this experiment. The pattern of the result for precision, recall and f-measure increased consistently as the word increased by 100. The value of precision corrected word stemmed for 100 words is 0.85, 200 words is 0.91, 300 words is 0.94, 400 words is 0.95 and 500 words is 0.95. Meanwhile, the recall values for each 100 words are 0.92, 0.94, 0.97, 0.97 and 0.97.

Table 4. Experimental stemming result for Precision, Recall, F-measure.

Total Word	**Precision**	**Recall**	**F-Measure**
100	0.85	0.92	0.84
200	0.91	0.94	0.92
300	0.94	0.97	0.95
400	0.95	0.97	0.96
500	0.95	0.97	0.97

The illustration result of stemming performance between words is shown in the histogram chart of Figure 6. From the result, the algorithm is able to stem words correctly as the word increases. This is because many words begin with the letter ‘P’ which is the affix word found from the word. The trend of precision consistently increases between the number of words experimented. When the word reaches 400 and 500 words, the value for precision is maintained which means the affix word found in these words is almost the same. Meanwhile, the value of recall becomes constant when the word reaches 300 data. The f-measure of the result is increasing as the word increases and the value is always between the precision and recall. This result shows the efficiency of the rule-based algorithm used in this study. A larger number of words used will increase the precision and accuracy.

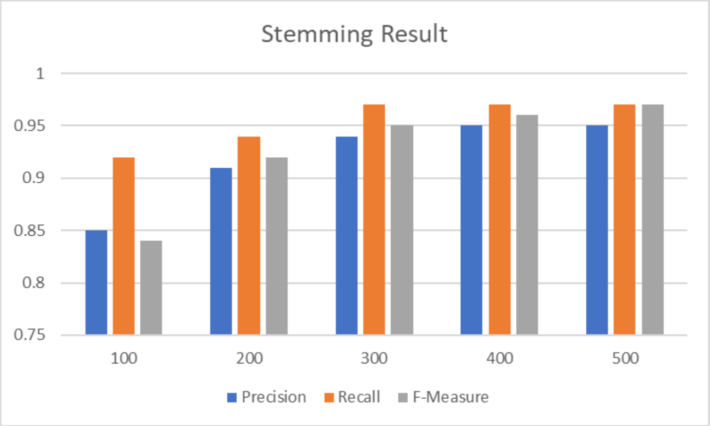

Figure 6. Comparison stemming performance between words.

CONCLUSION

This paper has discussed the stemming process using rule-based algorithm by obtaining affix word that starts with letter “p” from e-khutbah text. The stemming process focuses on removing prefix, suffix, including prefix and suffix for 500 words from 3233 sentences used for the experiment in this study. The experimental result obtained is significant. The algorithm used in this study can stem the words correctly. The assessment of the result is done by comparing the performance of precision, recall and f-measure for different size of word. As the number of words increases, the trend shows that the performance also increases. However, stemming result found that when the word reaches 400 and above, the precision result is maintained. This is because the affix word for 400 words and above is similar. The result of recall maintains when it reaches 300 words and above. Meanwhile, the f-measure result shows a slight and consistent increase as the number of words increases. Some limitations are found in this study that need further improvement such as conducting experiment on other types of affix word in the Malay language with larger words.

ACKNOWLEDGEMENT

This work is financially support by Institute of Research Management and Innovation, Universiti Teknologi MARA for the financial support in this research under the grant 600-RMC/LESTARI SDG-T 5/3 (145/2019).

REFERENCES

[1] Rifai, Wafda, “Modification of Stemming Algorithm Using A Non Deterministic Approach To Indonesian Text,” Indonesian *Journal of Computing and Cybernetics Systems*,13. 379,2019, doi: 10.22146/ijccs.49072.

[2] M. N. Kassim, M. A. Maarof, A. Zainal and A. A. Wahab,”Word stemming challenges in Malay texts: A literature review” 2016 4th International Conference on Information and Communication Technology (ICoICT), 2016, pp. 1-6, doi: 10.1109/ICoICT.2016.7571887.

[3] Boukhalfa, I., Mostefai, S., & Chekkai, N. , “A Study of Graph Based Stemmer in Arabic Extrinsic Plagiarism Detection,” Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, pp. 27-32, 2018.

[4] Permana, Y., & Emarilis, A., “Stemming Analysis Indonesian Language News Text with Porter Algorithm,” Journal of Physics: Conference Series, Vol. 1845, No. 1, p. 012019, IOP Publishing, 2021.

[5] Maheswari, S., & Arthi, K., “Rule Based Morphological Variation Removable Stemming Algorithm”, *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN, 2277-3878, 2019.

[6] Samuel, J., & Teferra, S.,“Designing A Rule Based Stemming Algorithm for Kambaata Language Text”, *International Journal of Computational Linguistics (IJCL)*, Volume 9 : Issue 2, 2018.

[7] Siswandi, Arif & Permana, Yudi & Emarilis, Arvita. , “Stemming Analysis Indonesian Language News Text with Porter Algorithm,” Journal of Physics: Conference Series. 1845. 012019, 2021, doi: 10.1088/1742-6596/1845/1/012019.

[8] Razmi, N. A, Zamri, M. Z., Ghazalli, S. S. S., & Seman, N., “Visualizing Stemming Techniques on Online News Articles Text Analytics”, Bulletin of Electrical Engineering and Informatics, [S.l.], v. 10, n. 1, p. 365-373, feb. 2021. ISSN 2302-9285. doi:https://doi.org/10.11591/eei.v10i1.2504.

[9] Khan, R. U., Mohamad, F. H, UlHaq, M. I., Adruce, S. A. Z., Anding, P. N., Khan, S. N., Al-Hababi, A. Y. S., “Malay Language Stemmer” *International Journal for Research In Emerging Science And Technology*, Volume 4: Issue 12,2019.