Implementation of Porter Stemmer Algorithm to Obtain Basic Words in Toba Batak Language Documents with the Two-Level Morphological Method

M A Muchtar¹, Erna B Nababan¹, Manguji Nababan², UAndayani¹, Tuti Simanjuntak¹, Opim S Sitompul¹

¹Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
²Pusat Dokumentasi dan Pengkajian Kebudayaan Batak, Universitas HKBP Nommensen, Medan, Indonesia

*Correspondence Email: anggi.muchtar@usu.ac.id

Abstract. Stemming is the process of deciphering words by carrying out affixes to words to get the right base words. In its development stemming has been developed for various languages, including Porter Stemmer. In this study, the stem was used to add Toba Batak language. Decapitation of affixes in the Toba Batak language is difficult because there are essential words that contain affixes and there are essential words that decay when affixed, namely the essential words beginning with the letter's 't', 's' and 'p'. To solve this problem by offering the implementation of a porter stemmer algorithm using the two-level morphological method. This stemmer method uses a dictionary of essential words. Each step in the stemmer process will always check the lexicon of the essential words so that the essential words that contain affixes will not be monitored, and if the word is not in the dictionary the basic word will not be stemming so when there are a name people or place names that contain affixes will not be implemented. The use of two levels of morphology was applied to the essential words beginning with the letter's 't', 's' and 'p' to get the right essential words. From the results of testing on 20 Toba Batak language documents, the Porter stemmer algorithm with two levels of morphology can produce an accuracy rate of 87.503% with an average speed of 0.000460798 seconds per word.

1. Introduction
Indonesia is a nation known for having many tribes. Each tribe has a different language and culture. For example, the Batak tribe. The Batak people themselves still consist of five sub-groups, namely Angkola-Mandailing, Toba, Simalungun, Pakpak-Dairi, and Karo [1]. The five Batak tribes have different languages as well, so the authors only focus on the Toba Batak language. The Toba Batak language is included in an ancient Malay language family that has its grammar system and the meaning of the word [2]. The Toba Batak language and Indonesian language come from the same family so that Indonesian grammar can be used as a reference tool. However, in the Toba Batak language, some words are not equivalent to Indonesian so that they need to be translated themselves. The Toba Batak language also has its phonetics, so the writing and pronunciation are different [3].

Stemming is a process for determining the essential words of a word by removing all affixes which consist of prefixes (prefixes), insertions (infixes), suffixes (suffixes) and combinations of prefixes and
suffixes (confixes) in derivative words. One of the algorithms used for stemming is the Porter stemmer algorithm. The morphology of the Indonesian language is not much different from the morphology of the Toba Batak language, which consists of prefixes, inserts, suffixes, proprietary pronouns, and particles [3]. The affix syllabification in the Toba Batak language are difficult because there is an affix which turns out to be part of the essential words themselves. For example, the word "manakko' mencuri'. If the word is stemmed with one level of morphology, the basic word is "akko" which should be "takko". In order to get an accurate root word, it must be solved with two levels of morphology.

In this study, the system to be built is a system for determining essential words in Toba Batak language documents. The input of this system is the text of the Toba Batak language in a document that will be processed using the Porter stemmer algorithm to produce output in the form of input documents before the system and after the order. Words that cannot be applied using a single morphological level will be used using two levels of morphology to obtain more accurate results [4].

Previous studies that have been conducted include [5], the approach taken is two levels of morphology that are implemented with xfst and lexc. This approach has been able to handle non-concatenative reduplication and morphology.

Subsequent research [6], modifications made to the rules of the stemmer porter by combining them with the regulations on Nazief & Adriani’s algorithm can improve the accuracy of the stem. The complete dictionary used for stemming processes affects the accuracy of the stem. Increasing the efficiency of the stem in many cases of search decreases the value of precision but does not affect the recall value.

Subsequent research using the Porter stemmer algorithm is [4]. By applying the 2-level morphology rule, it can increase the ability of stemmer to get the right base words on the base words beginning with the letter's "k", "t", "s" and "p" with 100% accuracy. The prefix and suffix combination rules can also increase the stemmer's ability to determine the suffix attached to a word especially for the "-kan" and "-a" suffixes with an accuracy rate of 96.6%. But the phrase homonym, homophone, homograph, and polysemy still cause problems. Besides, a method is also needed to identify the base words on repeated words correctly.

Furthermore, [7] with an accuracy level of 0.66 while the average processing time for each word is 0.000000160 seconds. The results of stemming accuracy are very dependent on the collection of original word dictionaries related to affixes — the more complete the group, the better the accuracy of the stemming results.

Next is [8], where the Porter stemmer algorithm determines the essential words in the inputted document and helps speed up the winnowing algorithm in determining the fingerprint value of a text so that the results are in the form of similarity between two coded documents.

The next study [9] is the utilization of the Porter stemmer algorithm for Indonesian in the process of classifying book types. The text pre-processing stage that is carried out on the tested document documents makes it easier for the stemming process. Stemming algorithms with stemmer porters for Indonesian that use rule-based algorithms significantly help speed up the classification process because the word generated from the test documents is the proper base word to facilitate matching with keywords and calculating the probability of occurrence. The prototype built with the application of porter stemmer rules for Indonesian can be used to help speed up and simplify the process of classifying book types. Other study shows how to do the next stage of text processing by summarizing the paragraph [10] which can be applied too later on in this Batak documents.

2. Methodology
The method proposed for determining the basics in the Toba Batak language document consists of several steps. The steps are the collection of Toba Batak language documents that will be used as test data; tokenization is to divide sentence by sentence into words per word (token) and syllabification (stemming). After the steps are taken, you will get the essential words in the document. Each of these stages will be explained in more detail in the next section. The general architecture that describes each step of the method used in this study is shown in Figure 1.
The description are as follows:

a. **Input**
   
The input used in this system is a text file for Toba Batak Latin language documents with an extension of .docx. The document is inputted to be processed by the system.

b. **Preprocessing**
   
   1. Filtering
      
      Filtering is the process of removing punctuation marks and characters that are considered invalid, such as numbers and symbols. For example, the sentence "You are on Sipahutar taon 1995." will be filtered into "You are on Sipahutar taon".

   2. Case Folding
      
      Case-folding is the process of converting an entire text of a document into a standard form, namely lowercase letters. The purpose of this case-folding is to equalize the case in the document because in a document it is not all capital letters or all lowercase letters. For example, from the results of filtering, "You always on Sipahutar taon". Be "au on the taon line".

   3. Tokenisation
      
      Tokenisation is the process of cutting a document into parts called tokens. After the filtering process and folding case are completed, the tokenization will be carried out to produce words per word that have been separated. For example, the sentence "Halaki do na manakko abit i." Is typified into [halaki] [do] [na] [manakko] [abit] [i].
c. **Process Stemming**

The steps taken in the stemming process are as follows.

1. Check the dictionary, if the word is in the dictionary, then the word is a basic word, and if it is not there then going to stage 2.
2. Check prefix 1 (ma-, mang-, mar-, pan-, par-, tar-), if there is no step 4, if there is the first prefix, then check the dictionary. If the word is in the dictionary, then the word is the original word, if it does not go to stage 3.
3. Apply two levels of morphology then check the glossary, if any, the word is the base word, if it does not go to stage 4.
4. Check the second prefix (si-, na-, ni-, di-, hu-, ha-, tu-, um-, pasi-, hina-), if there is no continue to stage 5, if there is delete prefix 2 then check the dictionary, if the word is in the dictionary, then the word is the primary word, if it does not go to step 5.
5. Check the particle, if there is no further step 6 if there is delete the particle then check the dictionary if the word is in the dictionary then the word is the essential word. If there is no further, go to step 6.
6. Check the proprietary pronouns, if there is no progress to stage 7, if there are delete pronouns then check the dictionary. If the word is in the dictionary, the word is the essential word. If there is no continuing to stage 7.
7. Check suffix, if it does not exist, the word is considered as the base word and returns the original word. If there is, delete the suffix then check the dictionary. If the word is in the dictionary, the word is a base word; if it does not exist, the word is considered as the base word and returns the original word.

Examples of sentences from the results of the previous process, namely [halaki] [manakko] [abit] [i] will be changed to [halak] [i] [man] [takko] [abit] [i]. The word "halaki" goes through process 1-6 and enters the 7th process which is removing the ending "-i". The word "manakko" goes through the 1-2 process and enters the third process, namely removing the "man-" prefix by applying two levels of morphology. While the word "abit" only enters the first process.

d. **Output**

The output of the design of this system is stemming results in the form of essential words from the inputted document. After the stemming process is complete, the stemming results in the form of essential words can be saved with extension of Microsoft Word documents.

3. **Results and Discussion**

When the system is completed, the order will be tested. The testing of this system aims to determine whether the system is functioning correctly and measure the accuracy of the Porter stemmer algorithm with the two-level morphological method implemented in the Toba Batak language stemming system. The accuracy of the results of testing this system will be calculated using the following formula (1).

\[
\text{Accuracy} = \frac{\text{Total Successful Stemmed Words}}{\text{Total Stemmed Words}} \times 100\%
\]

The results of system testing can be seen in Table 1.
Table 1. Result of a testing system.

| Input Document | Amount of words in Document | Amount of Stemmed words | Amount of Successful Stemmed Words | Accurate (%) |
|-----------------|-----------------------------|-------------------------|-----------------------------------|--------------|
| Doc 1           | 268                         | 60                      | 56                                | 93,33        |
| Doc 2           | 288                         | 59                      | 56                                | 94,91        |
| Doc 3           | 378                         | 60                      | 50                                | 83,33        |
| Doc 4           | 386                         | 81                      | 71                                | 87,65        |
| Doc 5           | 175                         | 43                      | 39                                | 90,69        |
| Doc 6           | 232                         | 60                      | 54                                | 93,33        |
| Doc 7           | 777                         | 208                     | 191                               | 91,82        |
| Doc 8           | 209                         | 52                      | 47                                | 90,38        |
| Doc 9           | 663                         | 181                     | 159                               | 87,84        |
| Doc 10          | 846                         | 225                     | 188                               | 73,72        |
| Doc 11          | 230                         | 83                      | 79                                | 95,18        |
| Doc 12          | 208                         | 78                      | 74                                | 94,87        |
| Doc 13          | 156                         | 77                      | 70                                | 90,90        |
| Doc 14          | 349                         | 88                      | 77                                | 87,5         |
| Doc 15          | 252                         | 70                      | 57                                | 81,42        |
| Doc 16          | 205                         | 79                      | 72                                | 91,13        |
| Doc 17          | 218                         | 63                      | 58                                | 92,06        |
| Doc 18          | 196                         | 71                      | 67                                | 94,36        |
| Doc 19          | 211                         | 65                      | 62                                | 95,38        |
| Doc 20          | 241                         | 76                      | 61                                | 80,26        |

The Average of Accuration 89,503

Based on the results of testing, as found in Table 1, several affixes cannot be implemented by the Porter Stem algorithm with the two-level morphological method. The stemming failure is caused by several things, such as those found in Table 2.

Table 2. Stemming error.

| No. | Error                                           | Example                                           | Description                                                                 |
|-----|-------------------------------------------------|---------------------------------------------------|-----------------------------------------------------------------------------|
| 1   | The basic word contains more than two prefixes  | Mar-si-pa-sangap-an, Mar-si-pa-sombu-an, Mar-si-pa-ture-an, Mar-na-man-sohot, etc. | The porter stemmer algorithm used in this study can only delete one and two prefixes |
| 2   | Essential words contain more than one suffix    | Ihut-hon-on, Togi-hon-on, Hudus-hon-on, etc.      | The porter stemmer algorithm used in this study can only delete one suffix   |
| 3   | Essential words ending in vowels and affixed with pronouns belonging to hu and ta | Rohanghu, Rohanta, Pisonghu, Pisonta, etc.       | The essential words ending in vowels and spelled with the pronouns belonging to -hu and -ta there are additional letters ‘ng ‘and ‘n’. |
| 4   | Essential words containing prefixes and affixed endings, | Panganan (pangan), Pamatangna (pamatang), Haemuna (hae), | The first process carried out in the Porter stemmer algorithm used in this study is to delete the prefix so |
4. Conclusion
There are several conclusions:
- The porter stemmer algorithm with two morphological levels can be applied to the Toba Batak language with an accuracy rate of 89.503%.
- By using two standards of morphology to the essential words that decay if given a prefix, namely the critical words are beginning with the letter's 't', 's' and 'p', get the right essential words.
- Words that are systemized but the results of stemming are not in the dictionary will be returned to the original word. For example, a person's name and place name that is not in the dictionary.
- The more complete the word data in the dictionary, the stemming results will be more accurate because if the word is not in the dictionary, then the word will not be programmed.

Further research:
The development of further research are as follows:
1. It is hoped that further research can overcome the shortcomings of this stemming system, such as:
   a. To be able to be stemming on affixed words that have more than two prefixes and more than one suffix.
   b. To be able to be stemming on the essential words ending in vowels and spiked with pronouns belonging to hu
   c. To be able to be stemming on essential words that contain prefixes and affixed suffixes, particles, and proprietary pronouns.
   d. To be able to do stemming on crucial words that are attached with inserts
2. The porter stemmer algorithm can be combined with other stemming algorithms to improve stemming accuracy results.
3. Stemming results can be used to do translations into other languages
4. This Toba Batak language stemming system can be developed in the mobile version.

Acknowledgments
This research was supported by the Universitas Sumatera Utara. The study program of Information Technology Lecturer and Staff member. Mr. Manguji Nababan as lecturer and Pusat Dokumentasi Pengkajian Kebudayaan Batak (PD & PKB) Universitas Nommensen Medan who was willing to provide research data to the Author.

References
[1] Sibarani, Robert. 1997. Sintaksis Bahasa Batak Toba. Medan: USU Press.
[2] Malau, G. G. 2000. Aneka Ragam Budaya Batak. Yayasan Bina Nusantara: Jakarta.
[3] Sinaga, A. B. 2002. Tata Bahasa Batak Toba: Meresapkan Jiwa dan Darah Batak. Bina Media: Medan.
[4] Wiguna, P. B. S. & Bimo, S. H. 2013. Peningkatan Algoritma Porter Stemmer Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran. *JNTETI*, Vol. 2, No. 2.

[5] Pisceldo, F., Mahendra, R. & Manurung, R. 2008. A Two-Level Morphological Analyser for The Indonesian Language. *Australian Language Technology Association Workshop*, Vol. 6.

[6] Budiasih, N. N., Wirayuda, T. A. B. & Dayawati, R. N. 2009. Analisis Dan Implementasi Stemming Teks Berbahasa Indonesia Dengan Menggunakan Porter Stemmer. Tugas Akhir Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom: Bandung.

[7] Zaman, B. 2014. Modifikasi Algoritma Porter Untuk Stemming Pada Kata Bahasa Indonesia. *Seminar Nasional Teknologi Informasi dan Komunikasi*: Yogyakarta.

[8] Ferari, L.A. 2014. Implementasi Algoritma Winnowing dan Porter Stemmer Mendeteksi Kemiripan Dua Dokumen Berbasis Web. Tugas Akhir Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara: Medan.

[9] Indriyono, B. V., Utami, E. & Sunyoto, A. 2015. Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku. *Jurnal Buana Informatika*, Vol. 6, No. 4.

[10] Gunawan D, Pasaribu A, Rahmat R F, Budiarto R 2017 Automatic Text Summarization for Indonesian Language using TextTeaser *IOP Conference Series : Material Science and Engineering* *109*(1) 012048