Stemming Analysis Indonesian Language News Text with Porter Algorithm

Arif Siswandi¹, A. Yudi Permana², Arvita Emarilis³

Universitas Pelita Bangsa, Bekasi, Indonesia

Email: arif.siswandi@pelitabangsa.ac.id¹, yudi@pelitabangsa.ac.id², arvita@pelitabangsa.ac.id³

Abstract. Stemming is the process of classifying various morphological variations of a word or sentence into one and the same basic form. In Indonesian language stemming, there are two types of stemming methods that already exist, namely the dictionary-based stemming algorithm and the non-dictionary-based stemming algorithm. In this study the algorithm used is the Indonesian Porter algorithm for dictionary-based ones. The test was carried out using 100 predetermined Indonesian text documents. The results of tests conducted show that the highest accuracy value is found in the Porter algorithm, the least Overstemming and Understemming values are also found in the Porter algorithm.

Keywords: Information Retrieval, Stemming, Accuracy, Overstemming and Understemming.

1. Introduction

In conducting information search of a textual documents or otherwise known as Information Retrieval (IR) is a process of separating documents deemed relevant from a collection of available documents. With an increase in the number of accessible text documents on the Internet followed by an increased user’s need for search devices and information that is both effective and efficient [1]. Effective means users get relevant documents to a query entered. Efficient is also a shorter search time result.

Stemming is a process that maps variant forms of words into root words [2]. Stemming is the essence of an information retrieval that is effective and efficient and is widely accepted by user. Stemming is a process of finding the root word of a word. By omits all good prefixes, infixes, suffixes and confixes (combinations of prefixes and suffixes) in the derivative word. Stemming can also be used to support the process of categorization or classification and clustering. Stemming is used to replace the shape or configuration of a word to a root word and of that word that has aligned with the correct and proper morphological structure of Indonesia language.

The use of stemming in the Indonesian language includes two known types of stemming methods, i.e. stemming based dictionaries (dictionary based) and stemming based non-dictionaries (purely rule based). In the stemming algorithm which does not use a dictionary has a relatively high margin of error, but on the other hand the algorithm has an advantage in shorter process times than stemming algorithms based on a dictionary. Most stemming algorithms still rely on or on dictionaries to check whether the root word of a sentence or the word that has been done the stemming process is either discovered or not. When the root word is successfully found in the dictionary, the stemming process is suspended.

In stemming in Indonesian, there are several approaches such as stemming Porter, Tala,Vega,
Arifin and Setiono, Nazief and Adriani. There is hardly any general consensus on the effectiveness of the techniques of that approach. Another problem of stemming is still dependent from several such stemming techniques on a broader dictionary (comprehensive dictionary). In the study will be conducted effective measurements of the algorithm used in stemming that the dictionary uses the porter algorithm. Assessments of measurements are conducted on an accurate level, overstemming and understemming.

2. Stemming Porter Algorithm

Stemming specifically English discovered by Martin Porter in 1980. An algorithm’s search for a root word repeats by throwing the prefix (or rather the suffixes) at word-english because in English there is no prefix. Since English came from different classes, some modifications have been made to create porter algorithms so that they can be used according to Indonesian. Porter Stemmer for Indonesian based on English Porter Stemmer developed by W.B. Frakes in 1992.

![Stemming Porter Indonesian Process Flow](image)

There are five sets of rules on Porter’s Algorithm for Indonesia Language [1]. The rules can be seen at Tables 1 through Tables 5.

| Table 1. Rules for Inflectional Particle |
|-----------------------------------------|
| Suffixes | Replacement | Measure Condition | Additional Condition | Example |
|----------|-------------|--------------------|----------------------|---------|
| kah      | Null        | 2                  | Null                 | sakitkah → sakit |
| lah      | Null        | 2                  | Null                 | adalah → ada   |
| pun      | Null        | 2                  | Null                 | jalanpun → jalan|

| Table 2. Rules for Inflectional Kata Ganti Kepunyaan |
|------------------------------------------------------|
| Suffixes | Replacement | Measure Condition | Additional Condition | Example  |
|----------|-------------|--------------------|----------------------|---------|
| ku       | Null        | 2                  | Null                 | tanganku → tangan |
| mu       | Null        | 2                  | Null                 | tangannya → tangan|
| nya      | Null        | 2                  | Null                 |                     |

| Table 3. Rules for First Order Derivational Prefix |
|-----------------------------------------------------|
| Prefixes  | Replacement | Measure Condition | Additional Condition | Example    |
|-----------|-------------|--------------------|----------------------|------------|
| meng      | Null        | 2                  | Null                 | mengambil → ambil |
Table 4. Rules for Second Order Derivational Prefix

| Prefixes | Replacement | Measure Condition | Additional Condition | Example |
|----------|-------------|-------------------|----------------------|---------|
| meny     | s           | 2                 | V…*                 | menyapa → sapa |
| men      | Null        | 2                 | Null                 | mendapat → dapat |
| mem      | p           | 2                 | V…                  | memilih → pilih |
| mem      | Null        | 2                 | Null                 | membeli → beli |
| me       | Null        | 2                 | Null                 | merusak → rusak |
| peng     | Null        | 2                 | Null                 | penguji → uji |
| peny     | s           | 2                 | V…                  | penyayang → sayang |
| pen      | Null        | 2                 | Null                 | penduga → duga |
| pem      | p           | 2                 | V…                  | pemikir → pikir |
| pem      | Null        | 2                 | Null                 | pembaca → baca |
| di       | Null        | 2                 | Null                 | diuji → uji |
| ter      | Null        | 2                 | Null                 | tersapu → sapu |
| ke       | Null        | 2                 | Null                 | kekasih → kasih |

Table 5. Rules for Derivational Suffix

| Suffixes | Replacement | Measure Condition | Additional Condition | Example |
|----------|-------------|-------------------|----------------------|---------|
| kan      | Null        | 2                 | Prefix ≠ {ke,peng}  | tarikkan → tarik |
|           |             |                   | (meng)ambilkan → ambil |
| an       | Null        | 2                 | Prefix ≠ {di,meng,ter} | minuman → minum (per)janjian → janji |
|           |             |                   |                      | (men)dapati → dapat |
| i        | Null        | 2                 | Prefix ≠ {ber,ke,peng} | tandai → tanda |

Fig. 2. Flowchart of stemming Porter process of Indonesia Language
3. Research Methodology

Information on the methods of research that are carried out in this study. Formasi tentang metode penelitian yang dilakukan dalam penelitian ini. First from research objects, research design and data collection techniques performed from multiple sources, then doing data preprocessing and algorithm applications. The next step that will be taken is to experiment or test data set on each of the algorithms used. The flow process of Stemming Porter Indonesia language is reflected in the following stemming process images.

![Flowchart of Stemming Porter process of Indonesia Language](image)

**Fig. 3.** Flowchart of Stemming Porter process of Indonesia Language

4. Result

The discussion will be focused on trial results from the collection of documents used, a primary lexicon, and judgment to compare the root words stemming from the root words stemming from the root words according to human knowledge. Based on the research that has been done on getting different results between one algorithm and another. These tests are run 100 times each for each algorithm. To know the rate of performance of each algorithm would be measuring the value of the accuracy of words, the words overstemming and understemming.

4.1. Document Collection

For the collection of documents used in testing, there are sample documents of 100 documents, which have already been made using extenions .txt. The words in this document are obtained from the content of articles and news, both news and articles on engineering, health, science and electronic media. The total of words on 100 documents is 25,819 words. The contents of the words in each document vary from dozens of words to thousands of words, where the words in the document have not done the processing or preprocessing.

4.2. The Dictionary

More complete the basic vocabulary is used, the greater value of stemming accuracy. In the study, a baseline dictionary is drawn from the default list of main words in the Indonesian dictionary (KBBI) luring CHM V1.5 downloaded from ebsoft.web.id. The total number of root words in the dictionary is 31,295 root words [4].
4.3. Relevance Judgments
Our knowledge as humans of good and true root words is needed, since the higher level of human knowledge of the root word as well as a habitual word produces has better results. This is done to compare the root word of stemming result using computers and the root word produced of human knowledge.

Table 6. Relevance judgments on documents

| No | Input Words | Stemming Result (Root Word) | Relevance Judgments (Root Word) |
|----|-------------|-----------------------------|---------------------------------|
| 1. | angkasa     | angkasa                     | angkasa                         |
| 2. | adalah      | adalah                      | adalah                          |
| 3. | atas        | atas                         | atas                            |
| 4. | atmosfer    | atmosfer                    | atmosfer                        |
| 5. | bulanan     | bulan                       | bulan                           |
| 6. | bumi        | bumi                        | bumi                            |
| 7. | dari        | dari                        | dari                            |
| 8. | lapisan     | lapis                       | lapis                           |
| 9. | gas         | gas                         | gas                             |
|10. | yang        | yang                        | yang                            |

4.4. Measuring Evaluations
The stemming algorithm will be tested using 100 documents. As for the measurements that were done are as follows:

a) Accurate testing results are obtained from comparative stemming results with relevance judgments, divided by the number of words in the document.

\[
\text{Accuracy} = \frac{\text{Root words of stemming result} - \text{Relevance judgments}}{\text{Number of words in the document}} \times 100\% \quad (1)
\]

b) Overstemming is the words that are many cut words after the stemming process is compared to relevance judgments.

c) Understemming is the words that are a bit cut words after a stemming process compared to relevance judgments.

4.5. Test Result
The stemming algorithm will be tested using 100 documents. As for the measurements done are as follows: test results done on the value of the stemming accuracy process show that the average value of accuracy is on stemming Porter algorithm.

Table 7. Average accurate stemming results

| Document Number (100) | Porter Algorithm | Average accuracy (%) |
|-----------------------|------------------|----------------------|
|                       |                  | 94,470               |

Test results conducted on the value of the overstemming process indicate that the average overstemming on stemming Porter algorithm.

Table 8. Average percentage of overstemming

| Document Number (100) | Porter Algorithm | Average of overstemming (%) |
|-----------------------|------------------|----------------------------|
|                       |                  | 4,541                       |

Test results conducted on the value of the understemming process indicate that the average understemming is on stemming Porter algorithm.
Table 9. Average percentage of understemming

| Document Number (100) | Porter Algorithm |
|----------------------|------------------|
|                      |                  |

In Porter’s algorithm is when the word is not found in the dictionary database and then is thought to be the root word. There is the stemming result error on Porter’s algorithm against the affixes.

Table 10. Stemming results errors on Porter’s algorithm.

| Example          | Stemming Result | Should be  |
|------------------|-----------------|------------|
| Asupan           | asupan          | Asup       |
| Bartahun         | bartahun        | Tahun      |
| Bekerjasama      | bekerjasama     | Kerjasama  |
| Beratnya         | rat             | Berat      |
| Berekpresi       | berekpresi      | Ekspresi   |
| Berlaku          | berla           | Laku       |
| Berolah          | bero            | Olah       |
| berpengalaman    | berpengalaman   | Alam       |
| Bersalah         | bersa            | Salah      |
| Bersekolah       | seko            | Sekolah    |
| bertanggungjawab | bertanggungjawab| Tanggungjawab |
| Bertanya         | berta           | Tanya      |
| Bertopologi      | bertopologi     | Topologi   |
| Bukanlah         | bu              | Bukan      |
| Dariku           | dari            | Dari       |
| Dengan           | dengan          | Dengan     |
| Diadakan         | adakan          | Ada        |
| Dianjurkan       | dianjurkan      | Anjur      |
| Diataati         | diataati        | Taat       |

References

[1] Agusta, Ledy (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. Konferensi Nasional Sistem dan Informatika. KNS&I09-036.

[2] Tala, Fadillah. Z. 2004. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands

[3] Kamus Besar Bahasa Indonesia (KBBI) luring CHM V1.5, ebsoft.web.id. diakses Rabu 25 November 2015.

[4] Larkey, L. S., Ballesteros, L., and Connell, M.E. 2002. Improving Stemming for Arabic Information Retrieval : Light Stemming and Co-occurrence Analysis. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 11-15, Tampere, Finland.

[5] Nazief, Bobby dan Mirna Adriani. 1996. “Confix-Striping: Approach to Stemming Algorithm for Bahasa Indonesia”. Fakulty of Computer Science University of Indonesia.

[6] Nugraha, Lusianto Marga. 2010. Analisis Penggunaan Algoritma Stemming Vega pada Information Retrieval System. Telkom University.Permana.

[7] Fanissa Shima, Ali Fauzi M., Adinugroho Sigit: Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.2018

[8] Hamzah Amir, Naniek : Opinion Classification Using Maximum Entropy and K-Means Clustering, IEEE 2016
[9] Permana, A. Yudi, Ismasari Ismasari, and M. Makmun Effendi. "Optimasi Stemming Porter KBBI dan Cross Validation Naïve Bayes untuk Klasifikasi Topik Soal UN Bahasa Indonesia." Jurnal Ilmiah KOMPUTASI 17.4 (2018): 357-368.

[10] Tan, A. 1999. Text Mining: The state of the art and the challenges, In Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD’99 workshop on Knowledge Discovery from Advanced Databases.