

## **PENGARUH *TEXT PREPROCESSING* PADA *CLUSTERING* DOKUMEN TEKS BERBAHASA INDONESIA**

**Milatina, Abdul Syukur, Catur Supriyanto**  
Pascasarjana Teknik Informatika Universitas Dian Nuswantoro

### ***ABSTRACT***

*Document clustering is the process of grouping documents into similar topics Vector Space Model (VSM) represents a set of documents in the form of term-document matrix, where each column represents a document and each line represents the term (word) contained in the document. This study intended to determine the effect of various text preprocessing on Indonesian document clustering. Stopword, stemming and n-grams as text processing was examined. Document clustering algorithm was used in this thesis is k-means clustering algorithm. The results showed that combination of text preprocessing stage will affect the clustering of text documents in Indonesian language. The results showed the level of accuracy at each stage of different combinations of text preprocessing. The highest accuracy (0.883) was achieved by combining stopwords and 3-gram as preprocessing document clustering.*

*Keyword : Clustering, Vector Space Model, Text Preprocessing, K-Means*

### **1. LATAR BELAKANG**

Jumlah dokumen teks yang ada di internet tumbuh dengan sangat pesat sejalan dengan perkembangan teknologi informasi yang semakin maju, maka proses penyimpanan dokumen secara digital juga berkembang pesat. Dokumen teks ini dapat berupa static page, dynamic page, file dokumen, email, forum online dan blog. Aliran informasi berita yang berupa teks diperbarui setiap harinya dalam jumlah yang besar sehingga proses pengelompokan dokumen (document clustering) menjadi sangat penting. Clustering dokumen adalah proses pengelompokan dokumen yang memiliki kesamaan topik [1]. Tujuan dari proses clustering ini membagi dokumen berdasarkan kesamaan, sehingga memudahkan dalam proses pencarian.

Clustering dokumen telah lama diterapkan untuk memudahkan pengguna dalam mencari dokumen. Penerapan clustering ini berstandar pada suatu hipotesis (clusterhypothesis) bahwa dokumen yang relevan akan cenderung berada pada cluster yang sama jika pada koleksi dokumen dilakukan clustering. Beberapa penelitian sebelumnya yang berkaitan dengan clustering yang telah dilakukan antara lain: penelitian yang dilakukan oleh [2] yang menyebutkan bahwa pengaruh stemming memberikan efek pada akurasi clustering dokumen berbahasa Arab. Sedangkan penelitian bidang IR untuk dokumen Bahasa Indonesia telah dilakukan oleh [3] yang meneliti efek dari stemming pada Bahasa Indonesia. Sedangkan penelitian di bidang IR yang meneliti pengaruh teks preprocessing untuk clustering dokumen Bahasa Indonesia masih jarang dilakukan. Hal ini mengingat secara umum penelitian tentang komputasi bahasa untuk dokumen Bahasa Indonesia juga masih sangat minim [4]. Penelitian sebelumnya telah dilakukan oleh [5] yang meneliti tentang pengaruh preprocessing teks pada deteksi plagiarism. Sedangkan pada penelitian ini dimaksudkan untuk mengetahui pengaruh teks preprocessing dan kombinasinya pada akurasi clustering dokumen teks berBahasa Indonesia yang di ukur dengan F-Measure.

### **2. RUMUSAN MASALAH**

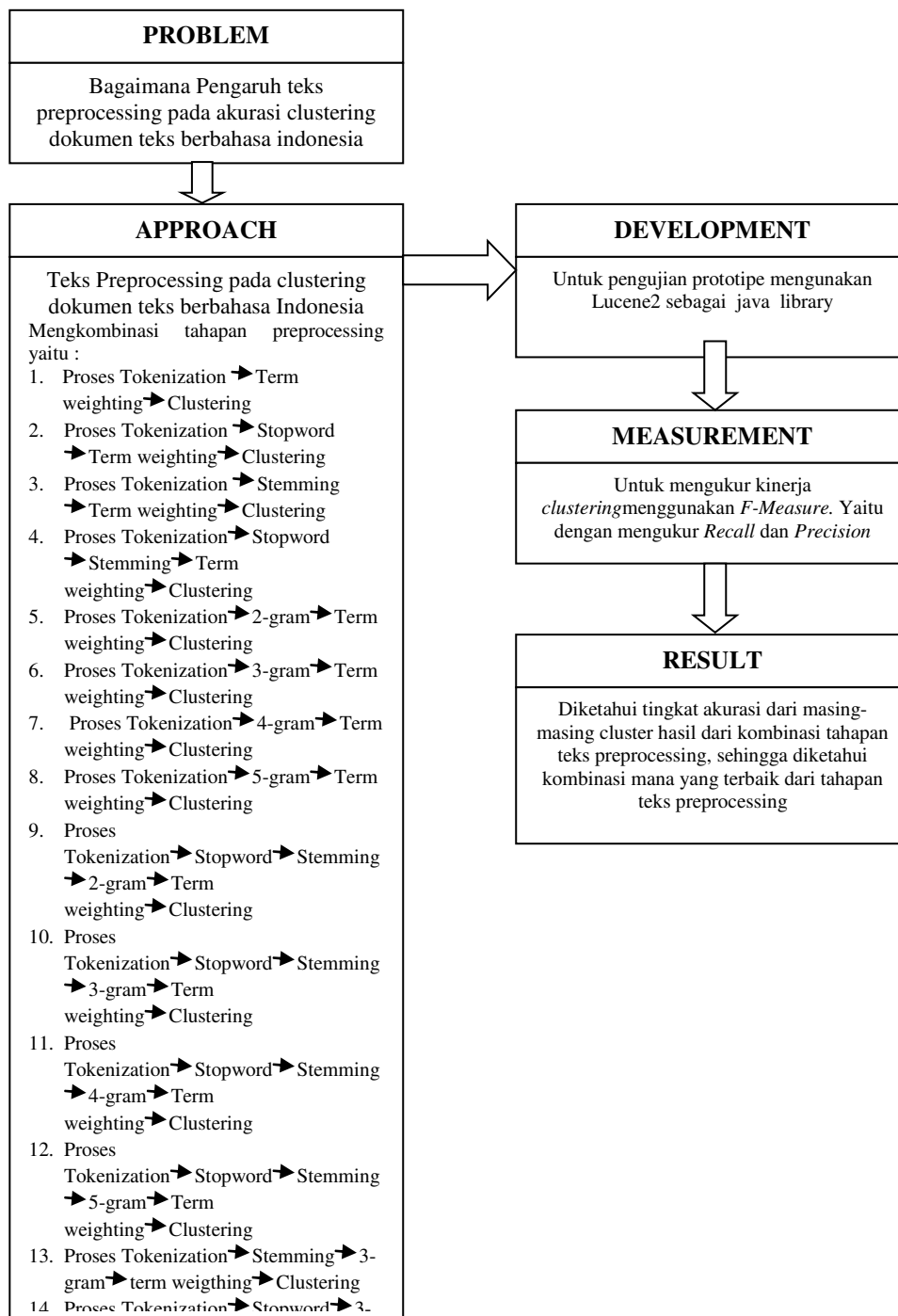
Belum diketahuinya pengaruh teks preprocessing pada akurasi dokumen clustering khususnya pada dokumen teks berBahasa Indonesia.

### 3. TUJUAN PENELITIAN

Tujuan penelitian ini adalah untuk:

1. Untuk mengetahui pengaruh dari teks preprocessing dan kombinasinya pada akurasi clustering dokumen teks berBahasa Indonesia.
2. Untuk mengetahui kombinasi terbaik dari teks preprocessing pada clustering dokumen teks berBahasa Indonesia.

### 4. KERANGKA PIKIR



## 5. LANDASAN TEORI

### 5.1. Preprocessing

Preprocessing adalah tahapan mengubah suatu dokumen kedalam format yang sesuai agar dapat diproses oleh algoritma clustering. Terdapat 3 tahapan preprocessing dalam penelitian ini, yaitu:

1. Tokenization, merupakan tahapan penguraian string teks menjadi term atau kata.
2. Stopword Removal, merupakan tahapan penghapusan kata-kata yang tidak relevan dalam penentuan topik sebuah dokumen dan yang sering muncul pada sebuah dokumen, misal "and", "or", "the", "a", "an" pada dokumen berbahasa inggris.
3. Stemming, merupakan tahapan pengubahan suatu kata menjadi akar kata nya dengan menghilangkan imbuhan awal atau akhir pada kata tersebut, misal kerjakan → kerja, bermain → main. Penelitian ini menggunakan algoritma porter stemmer.

### 5.2. Vector Space Model (VSM)

Vector Space Model (VSM) banyak digunakan dalam sistem temu kembali dokumen teks. VSM adalah model yang digunakan untuk mengukur kemiripan antar dokumen. VSM mengubah koleksi dokumen ke dalam matrik term-document [6]. Matrik term-document (Gambar 1) tersebut memiliki dimensi  $m \times n$  dimana  $m$  adalah jumlah term dan  $n$  adalah jumlah dokumen. Terdapat 3 metode pembobotan atau term weighting dalam VSM yaitu Term Frequency (TF), Invers Document Frequency (IDF) dan Term Frequency Invers Document Frequency (TFIDF). Gambar 1 menunjukkan Matriks Term Dokumen [7].

$$A_{max} = \begin{bmatrix} w_{11} & w_{12} & w_{1n} \\ \dots & \dots & \dots \\ w_{m1} & w_{m2} & w_{mn} \end{bmatrix}$$

↓      ↓      ↓

d1    d2    dn

← t1

← tm

Hasil tersebut masih dirasa belum dapat mengatasi kata dasar dengan baik, karena masih banyak kesalahan overstemming seperti pada contoh kata (sekolah) dan hasilnya menjadi (seko). Untuk mengatasi masalah yang dialami algoritma Confix Stripping, kemudian Putu Adhi Kerta mengembangkan algoritma ECS Stemmer (2008) [14 ], Revisi aturan dalam tabel aturan pemenggalan, Penggunaan loop, Pengembalian Akhiran di dalam algoritma stemmer yang ada pada algoritma Enhance Confix Stripping.

Hal ini dilakukan karena dirasa pada algoritma Enhance Confix Stripping masih terdapat kesalahan, terutama pada stemming kata serapan yang berasal dari bahasa asing. Kata serapan yang ada pada bahasa Indonesia ada yang berupa prefiks, infiks dan sufiks. Untuk mengatasi masalah tersebut, maka dibuatlah algoritma untuk mengatasi imbuhan (afiksasi) kata serapan asing yang masuk ke dalam bahasa Indonesia. Dalam penelitian ini ditujukan untuk membuat stemmer bahasa Indonesia yang dapat menyelesaikan kata serapan dalam bahasa Indonesia, sehingga pengukuran hasil stemming efektif disini menghitung tingkat kebenaran stemming, recall, precission, dan F-Measure pada algoritma stemming kata serapan pada bahasa Indonesia.

## 6. RUMUSAN MASALAH

Dari uraian latar belakang di atas maka masalah yang melatarbelakangi penelitian ini adalah belum terdapatnya stemming yang dapat mengatasi kata yang berafiksasi serapan asing.

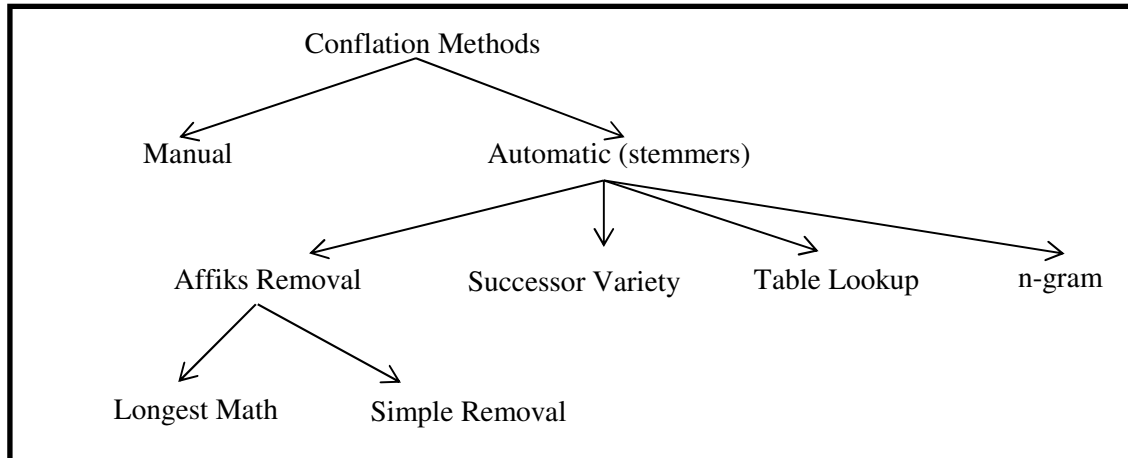
## 7. TUJUAN

Berdasarkan latar belakang dan rumusan masalah diatas, maka penelitian ini bertujuan untuk :

1. Melakukan modifikasi terhadap algoritma stemming agar dapat menstemming kata serapan bahasa Indonesia.
2. Mengukur recall, precision, accuracy,error rate dan F Measure hasil stemming kata serapan.

## 8. LANDASAN TEORI

Stemming adalah salah satu teknik untuk menyediakan cara untuk menemukan varian morfologi istilah pencarian.



Gambar Taksonomi untuk Algoritma Stemming

Kriteria untuk menilai keakuratan stemmers yaitu tidak terlalu banyak istilah yang dihapus (overstemming) dan tidak terlalu sedikit istilah yang dihapus (understemming). Efektifitas menemubalikkan informasi diukur dengan presisi, kecepatan proses, dan sebagainya.

Affiks Removal Stemmer dilakukan untuk menghilangkan awalan dan atau akhiran dari kata yang distem. Dalam banyak kasus affiks removal biasanya menggunakan Longest Match Stemmers yaitu sebuah iterasi dilakukan dengan penghilangan string terpanjang yang mungkin dari sebuah kata dengan mengacu pada kumpulan aturan tertentu. Proses ini akan diulangi sampai mendapatkan hasil yang diinginkan (kata dasar) atau tidak adanya substrings yang bisa dihapus lagi.

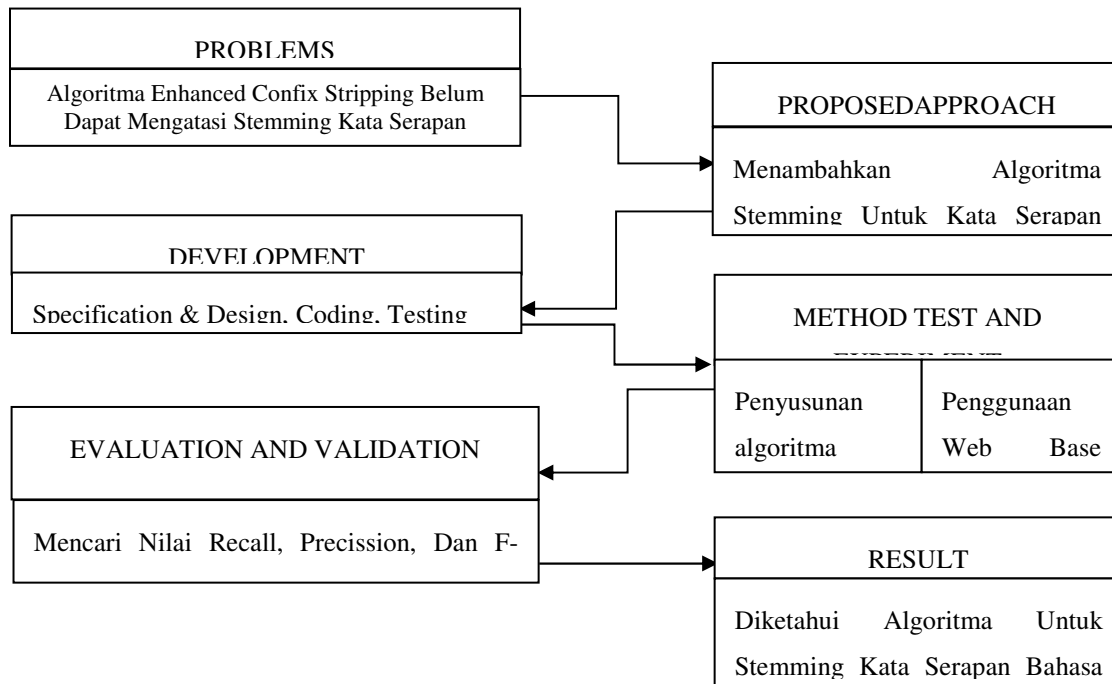
Successor Variety dikembangkan Hafer and Weiss 1974, didasarkan pada struktur bahasa untuk memisahkan kata dari dokumen dengan melihat distribusi dari fonem dalam suatu dokumen.

Table Lookup merupakan sebuah indeks tabel istilah yang disimpan, sehingga istilah dari query dan indeks bisa ditemukan dengan cepat.

N-gram dikembangkan oleh Adam dan Boreham (1974), menjelaskan tentang stemming menggunakan metoda diagram yang selanjutnya berkembang dengan sebutan N-Gram Methods. Metode ini menghitung persamaan term yang didasarkan pada jumlah unqi diagram yang dipakai bersama antar kata. Diagram merupakan substring yang diambil.

## 9. KERANGKA PEMIKIRAN

Secara umum metode penelitian yang telah dilaksanakan mengacu pada kerangka pemikiran sbb.



Gambar 5 Kerangka Pemikiran

## 10. HASIL PENELITIAN

Algoritma yang terdapat di Indonesia, khususnya yang digunakan untuk menangani pemenggalan kata dasar yang berimbuhan prefiks, konfiks, dan suffiks. Namun dalam prakteknya masih perlu dikembangkan algoritma stemming yang lebih baik recall dan presisi dalam menstemming kata berimbuhan dengan baik, dikarenakan beragamnya bahasa Indonesia yang banyak terpengaruh dari bahasa asing, sehingga sangat berpengaruh dalam pembentukan kata dalam bahasa Indonesia.

Algoritma stemming Enhanced Confix Stripping ternyata masih terdapat kelemahan yang mendasar dalam penstemmingan kata serapan yang masuk ke dalam bahasa Indonesia. Dibuktikan setelah dilakukan percobaan terhadap kata imbuhan serapan yang dimasukkan dalam stemmer ECS, tidak dapat distemming dengan benar. Hal ini dapat dilihat pada tabel dibawah ini.

Tabel 1 Sampel Hasil Kesalahan Stemming Algoritma Enhanced Confix Stripping

No	Kata	Tipe	Target Kata Dasar	Hasil ECS Stemmer	
				Hasil	Ket
1	Abnormalitas	Nomina	Abnormal	abnormalitas	Unstemming
2	Absenteisme	Nomina	Absenteis	absenteisme	Unstemming
3	Absolutisme	Nomina	Absolut	Absolutisme	Unstemming
4	Viskositas	Nomina	Viskos	Viskositas	Unstemming
5	Visualisasi	Nomina	Visual	Visualisasi	Unstemming

Dari data pada tabel 12 adalah sebagian, data selengkapnya terdapat pada halaman lampiran. Eksperimen yang dilakukan dalam penstemming kata serapan dalam bahasa Indonesia pada algoritma ECS sesuai tabel 12 didapatkan :

Jumlah kata yang distemming 1164

FN = 1021; TN = 79

FP = 64; TP = 0

Berdasarkan data yang didapat dari hasil eksperimen menggunakan ECS Stemmer untuk kata kata serapan bahasa Indonesia. Hasil dari presisi, recall dan akurasi dapat dilihat sebagai berikut:

Precision =  $TP / (TP + FP)$

=  $0 / (0 + 64)$

= 0

Recall =  $TP / (TP + FN)$

=  $0 / (0 + 1021)$

= 0

Akurasi =  $(TP + TN) / (TP + TN + FP + FN)$

=  $(0 + 79) / (0 + 79 + 64 + 1021)$

=  $79 / 1164$

= 0.067869416

Error Rate =  $(FP + FN) / (TP + TN + FP + FN)$

=  $(64 + 1021) / (0 + 79 + 64 + 1021)$

=  $1085 / 1164$

= 0.932130584

F measure =  $2 \frac{\text{presisi} * \text{recall}}{\text{presisi} + \text{recall}} = 2 \frac{0 * 0}{0 + 0}$

= 0

Keterangan:

TP (True Positive) = Kata Serapan yang memang tidak dapat distemming oleh ECS

FP (False Positive) = Kata nonserapan yang tidak dapat distemming oleh ECS, bisa jadi Understemming maupun Overstemming

TN (True Negative) = Kata nonserapan yang dapat di stemming oleh ECS

FN (False Negative) = Kata serapan yang tidak dapat distem oleh ECS

Analisa kesalahan stemming yang dilakukan oleh algoritma Enhance Confix Stripping tersebut diatas disebabkan karena langkah dari algoritma yang belum terdapat algoritma untuk menyelesaikan stemming kata serapan, rule algoritma

Kata yang belum di stemming dicari pada kamus. Jika kata itu langsung ditemukan, berarti kata tersebut adalah kata dasar. Kata tersebut dikembalikan dan algoritma dihentikan.

Hilangkan Inflectional suffixes terlebih dahulu. Jika hal ini berhasil dan suffix adalah partikel ("lah" atau "kah"), langkah ini dilakukan lagi untuk menghilangkan inflectional possessive pronoun suffixes ("ku", "mu" atau "nya").

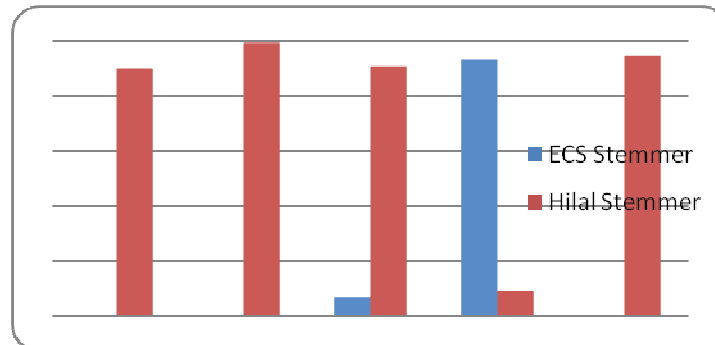
Derivational Suffix kemudian dihilangkan. Lalu langkah ini dilanjutkan lagi untuk mengecek apakah masih ada Derivational Suffix yang tersisa, jika ada maka dihilangkan. Jika tidak ada lagi maka lakukan langkah selanjutnya.

Kemudian Derivational Prefix dihilangkan. Lalu langkah ini dilanjutkan lagi untuk mengecek apakah masih ada Derivational Prefix yang tersisa, jika ada maka dihilangkan. Jika tidak ada lagi maka lakukan langkah selanjutnya.

Setelah tidak ada lagi imbuhan yang tersisa, maka algoritma ini dihentikan kemudian kata dasar tersebut dicari pada kamus, jika kata dasar tersebut ketemu berarti algoritma ini berhasil tapi jika kata dasar tersebut tidak ketemu pada kamus, maka dilakukan Recoding.

Jika semua langkah telah dilakukan tetapi kata dasar tersebut tidak ditemukan pada kamus juga maka algoritma ini mengembalikan kata yang asli sebelum dilakukan stemming.

Jika dilihat dari penjelasan rule diatas, dapat disimpulkan bahwa dalam algoritma ECS belum terdapat algoritma untuk menangani kata serapan yang masuk ke dalam bahasa Indonesia. Langkah yang dilakukan agar dapat menstemming kata berimbuhan serapan asing, yaitu menambahkan algoritma untuk membuang imbuhan kata serapan.



Gambar 6 Grafik hasil perbandingan dari Hilal Stemmer dengan ECS Stemmer

### 10.1. Pembahasan

Algoritma stemming untuk kata serapan bahasa Indonesia dilakukan untuk membuang suffiks –isasi, –logi, –itas, –ah, –at, –wan, –at, –wati, –wi, –in, –at, –isme, dan –me. Hal ini dilakukan karena banyaknya kata yang mendapat imbuhan sufiks tersebut dalam kata bahasa Indonesia. Selain sufiks tersebut, ditambahkan juga algoritma untuk menghapus infiks –in, –er, –el, dan ha. Untuk membuktikan algoritma untuk kata serapan bahasa Indonesia pada gambar 6 dan 7.

algoritma stemming untuk bahasa serapan bahasa Indonesia dilakukan beberapa perbaikan sebagai berikut :

Melakukan modifikasi dan penambahan aturan untuk kata yang berisuffiks – is, -isasi, -isme, -asi, -logi - wan, -wi, -iah, -at, -al, -ik, -if, -is, -logi, -or. dan -wati.

Melakukan penambahan aturan untuk menstemming kata yang berinfix –em, -in, -ha, -el.

Dari algoritma tersebut diatas dimaksudkan agar setiap kata yang berimbuhan afiksasi dari serapan bahasa lain dapat ditanggulangi dengan baik.

Penyelesaian kata serapan dapat dimaksimalkan jika ditambahkan rule yang digambarkan pada flowcart pada gambar 4.



2. Selanjutnya Hilangkan Inflectional suffixes terlebih dahulu. Jika hal ini berhasil dan suffix adalah partikel (“lah” atau ”kah”), langkah ini dilakukan lagi untuk menghilangkan inflectional possessive pronoun suffixes (“ku”, “mu” atau ”nya”). Jika ketemu kata dasar maka algoritma berhenti.
3. Jika terdapat kata berimbuhan Serapan asing – is, -isasi, -isme, -asi, -logi -wan, -wi, -iah, -at, -al, -ik, -is, -logi, -or. dan –wati, maka hapus suffiks serapan tersebut. Jika ditemukan kata dasar maka algoritma berhenti. Jika tidak berupa kata bersufiks serapan lanjut rule 4.
4. Derivational Suffix kemudian dihilangkan. Lalu langkah ini dilanjutkan lagi untuk mengecek apakah masih ada Derivational Suffix yang tersisa, jika ada maka dihilangkan. Jika tidak ada lagi maka lakukan langkah selanjutnya.
5. Kemudian Derivational Prefix dihilangkan. Lalu langkah ini dilanjutkan lagi untuk mengecek apakah masih ada Derivational Prefix yang tersisa, jika ada maka dihilangkan. Jika tidak ada lagi maka lakukan langkah selanjutnya.
6. Jika terdapat kata berinfiks –em, -in, -ha, dan -el, maka hapus infiks tersebut, jika ditemukan kata dasar, maka algoritma berhenti, jika tidak berupa kata yang berinfiks, lanjut ke rul berikutnya.
7. Setelah tidak ada lagi imbuhan yang tersisa, maka algoritma ini dihentikan kemudian kata dasar tersebut dicari pada kamus, jika kata dasar tersebut ketemu berarti algoritma ini berhasil tapi jika kata dasar tersebut tidak ketemu pada kamus, maka dilakukan Recoding.
8. Jika semua langkah telah dilakukan tetapi kata dasar tersebut tidak ditemukan pada kamus juga maka algoritma ini mengembalikan kata yang asli sebelum dilakukan stemming.

Analisa hasil eksperimen stemming kata serapan bahasa Indonesia yang mengalami overstemming dan Understemming dikarenakan masih sedikitnya data set kata dasar serapan bahasa Indonesia yang ada dalam kamus digital yang beredar saat ini. Kamus digital yang beredar di Indonesia bukan hanya kata dasar, tetapi masih terdapat kata yang tercampur dengan kata reduplikasi, serapan, dan lainnya. Hal inilah yang menjadikan kesalahan stemming kata serapan pada hasil eksperimen diatas. Unstemming dari hasil eksperimen pada keterangan ECS Stemmer terjadi dikarenakan ECS Stemmer belum terdapat algoritma yang digunakan untuk menstemming kata serapan. Jadi hasil dari eksperimen ditunjukkan Unstemming (Tidak dapat distemming) dan hasilnya tetap sama tanpa mengalami proses stemming.

Eksperimen yang dilakukan dalam penstemmingan kata serapan dalam bahasa Indonesia didapatkan

FN = 5; FP = 102  
TN = 140; TP = 917

Dari data diatas akan dijelaskan presisi, recall dan akurasi pada algoritma Stemming untuk kata serapan bahasa Indonesia. Hasil dari presisi, recall dan akurasi dapat dilihat sebagai berikut:

Precision =  $TP / (TP + FP)$   
=  $917 / (917 + 102)$   
=  $917 / 1019$   
= 0.89990186

Recall =  $TP / (TP + FN)$   
=  $917 / (917 + 5)$   
=  $917 / 922$   
= 0.99457701

Akurasi =  $(TP + TN) / (TP + TN + FP + FN)$   
=  $(917 + 140) / (917 + 140 + 102 + 5)$   
=  $1057 / 1164$   
= 0.90767903

Error Rate =  $(FP + FN) / (TP + TN + FP + FN)$   
=  $(102 + 5) / (917 + 140 + 102 + 5)$

$$\begin{aligned}
 &= 107 / 1159 \\
 &= 0.09232097 \\
 \text{F measure} &= 2 \frac{\text{presisi} * \text{recall}}{\text{presisi} + \text{recall}} = 2 \frac{0.89990186 * 0.99457701}{0.89990186 + 0.99457701} \\
 &= 0.94487378
 \end{aligned}$$

Keterangan:

TP (True Positive) = Kata Serapan yang dapat distemming oleh Hilal Stemming  
 FP (False Positive) = Kata nonserapan yang tidak dapat distemming oleh Hilal Stemming, bisa jadi Understemming maupun Overstemming  
 TN (True Negative) = Kata nonserapan yang dapat di stemming oleh Hilal Stemming  
 FN (False Negative) = Kata serapan yang tidak dapat distem oleh Hilal Stemming

## 11. KESIMPULAN

Dari hasil eksperimen algoritma Enhanced Confix Stripping dalam melakukan stemming kata serapan ditunjukkan nilai precision = 0, recall = 0, akurasi = 0.067869416, Error Rate = 0.932130584, F – Measure = ∞.

Hasil eksperimen algoritma stemming kata serapan menunjukkan keberhasilan stemming kata yang berimbuhan –isasi, –logi, –itas, –ah, –at, –wan, –at, –wati, –wi, –in, –at, –isme, dan –me. Selain suffiks kata serapan juga ditambahkan algoritma penghapusan infiks -em, -el, er, dan em dengan nilai precision = 0.89990186, recall = 0.99457701, Akurasi = 0.90767903, Error Rate = 0.09232097, dan F-Measure= 0.94487378. Algoritma stemming kata serapan bahasa indonesia dapat menangani stemming kata bersuffiks serapan asing dengan baik, tetapi waktu yang diperlukan untuk proses stemming lebih lama. Hal ini terjadi karena semakin bertambahnya rule stemming untuk mengatasi kata serapan dalam bahasa Indonesia.

## 12. SARAN

Dalam penelitian ini dikhususkan pada penambahan algoritma stemming untuk mengatasi kata serapan –isasi, –logi, –itas, –ah, –wan, –wati, –wi, –in, –at, –isme, dan –me dalam Bahasa Indonesia. Selain itu ditambahkan juga algoritma untuk mengatasi infiks em, el, er, dan ha. Dalam kenyataannya ternyata masih ditemukan masalah-masalah baru. yaitu pada kata serapan bahasa asing lainnya. Jika terdapat kata serapan lainnya dapat ditambahkan kedalam algoritma stemming karena pada prinsipnya hampir samadengan kata serapan yang ada pada penelitian ini.

Untuk mendapatkan nilai recall dan Precission serta akurasi yang lebih baik dalam algoritma stemming kata dasar serapan, masih perlu ditambahkan algoritma untuk mengatasi kata yang berprefiks dari serapan bahasa asing lainnya. Sehingga akan mampu menstemming kata serapan dengan lebih banyak ragamnya. Dalam penelitian ini masih menggunakan data set kata dasar serapan yang masih belum lengkap, untuk selanjutnya diharapkan dapat ditambahkan data set kamus kata dasar serapan yang lebih lengkap.

## 13. PENUTUP

Jurnal dengan judul “Algoritma Stemming Untuk Kata Serapan Bahasa Indonesia” ini dapat penulis selesaikan sesuai rencana karena dukungan dari berbagai pihak yang tidak ternilai besarnya.

## DAFTAR PUSTAKA

- [1] Fulayi Idi, "Building a French Stemmer Using a Dictionary Of French Root Words," University Putra Malaysia, Thesis 1999.

- [2] Ledy Agusta, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," Konferensi Nasional Sistem dan Informatika, 2009.
- [3] Adriani Mirna, Jelita Asian, Bobby Nazief, S.M.M. Tahaghoghi, and Hugh E. Williams, "Stemming Indonesian: A Confix-Stripping Approach," *ACM Transactions on Asian Language Information Processing*, Vol. 6, No. 4, Article 13, December 2007.
- [4] Jelita Asian, Mirna Adriani, and Bobby Nazief, "Stemming Indonesian: A Confix-Stripping Approach," *ACM Transactions on Asian Language Information Processing*, vol. 6, pp. 13-32, 2007.
- [5] Jelita Asian, Hugh E. Williams, and S.M.M. Tahaghoghi, "Stemming Indonesian," School of Computer Science and Information Technology RMIT University, GPO Box 2476V, Melbourne 3001, Australia.
- [6] Fadillah Z Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Universiteit van Amsterdam, Netherlands, Thesis 2003.
- [7] Pusat Bahasa, Tata Bahasa Baku Bahasa Indonesia. Republik of Indonesia: Balai Pustaka Dept. of Cultural and Education, 1988.
- [8] Dept. of Cultural and Education, Pedoman Umum Ejaan, editor, Ed. Republic of Indonesia: Pustaka Setia, 1987.
- [9] Sock Yin Tai, Cheng Soon Ong, and Noor Aida Abdullah, "On Designing an Automated Malaysian Stemmer for the Malay language," in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, 2000.
- [10] Lily Suryana Indradjaja and Stephane Bressan, "Automatic Learning of Stemming Rules for the Indonesian Language," pp. 55-62, 2003.
- [11] Georgios Ntais, *Development of a Stemmer for the Greek Language.*: Department of Computer and Systems Sciences, 2006.
- [12] Kumar Santosh M. and Kavi Narayana, "Corpus Based Statistical approaches for stemming telugu," *Journal of quantative linguistic*, vol. 16, no. 1, pp. 130-133, 2006.
- [13] Agus Zainal Arifin, I Putu Adhi Kerta Mahendra, and Henning Titi Ciptaningtyas, "Enhanced Confix Stripping Stemmer And Ants Algorithm For Classifying News Document In Indonesian Language," in *Proceeding of International Conference on Information & Communication Technology and Systems (ICTS)*, 2009.