STEMMING ARTIKEL BERBAHASA INDONESIA DENGAN PENDEKATAN *CONFIX-STRIPPING*

Rinci Kembang Hapsari dan Yunus Juli Santoso
Teknik Informatika, Institut Teknologi Adhi Tama Surabaya, Surabaya, Indonesia
e-mail: kembanghapsari@gmail.com

ABSTRAK

Dalam bahasa Indonesia terdapat kata dasar dan kata imbuhan. Sebuah kata berimbuhan dapat terbentuk dengan adanya penambahan imbuhan awalan, sisipan ataupun akhiran pada sebuah kata dasar. Sehingga sebuah kata bisa ditemukan kata dasarnya dengan cara menghilangkan atau menghapus imbuhan-imbuhan kata tersebut.
Stemming adalah proses pemetaan dan penguraian berbagai bentuk kata menjadi bentuk dasarnya. Proses pemetaan dan penguraian digunakan untuk menemukan kata dasar dari sebuah kata yang mengalami imbuhan dengan cara menghilangkan atau menghapus imbuhan-imbuhan tersebut.
Tujuan penelitian ini adalah mencari kata dasar dari kata imbuhan teks bebahasa Indonesia dalam sebuah dokumen. Teknik *Stemming* yang digunakan terdiri dari 3 tahap, yaitu tahap pertama, *parsing* dokumen dilakukan untuk memecah sebuah dokumen menjadi kata-kata atau disebut *token.* Tahap kedua, *stopword* yang merupakan proses penghilangan kata yang tidak penting dalam dokumen. Dan tahap ketiga, proses *confix-stripping.* Dimana *Confix-Stripping* merupakan algoritma *Stemming* untuk pemenggalan atau pengupasan kata imbuhan awalan, akhiran, dan gabungan antara awalan-akhiran menjadi bentuk dasar. Berdasarkan hasil pengujian yang dilakukan *Stemming* terhadap beberapa kata dalam dokumen dengan menggunakan pendekatan *Confix-Stripping*, dihasilkan rata-rata nilai akurasi yang diperoleh sebesar 94.85% dari 20 dokumen teks berbahasa Indonesia yang diuji.

Kata kunci: *Steamming, Parsing, Stopword, Confix-Stripping, Token.*

PENDAHULUAN

Bahasa Indonesia adalah bahasa aglutinatif yang memungkinkan kata-kata baru akan dibentuk dengan menambahkan prefix dan sufiks untuk kata (Quinn, 2001). Kata-kata baru juga dapat dibentuk dengan mengulang kata dan dengan memasukkan infiks menjadi sebuah kata. Paice menyatakan bahwa kata-kata biasanya *Stemming* karena bentuk yang berbeda sintaksis diasumsikan memiliki arti yang sama (Paice, 1994). *Stemming* membutuhkan pemahaman yang baik tentang bahasa yang bersangkutan. Seperti bahasa Inggris memiliki awalan "hiper-" dalam "hipertensi" dan "hiperaktif", "anti-" dalam "antisosial", dan "ultra-" dalam "ultraviolet". Awalan ini menciptakan makna baru yang berbeda dari makna aslinya, karena itu mereka tidak mempertimbangkan dalam *Information Retrieval* (*IR*).Dalam bahasa Inggris *Stemming* biasanya hanya menghapus akhiran.

Implementasi pada penelitian ini didasarkan pada Inggris Porter Stemmer dikembangkan oleh (Frakes, 1992).*Stemming* bahasa Inggris dan bahasa Indonesia

ISBN: 978-602-70604-1-8

berasal dari dua kelas bahasa yang berbeda, sehingga beberapa modifikasi harus dilakukan untuk membuat algoritma yang cocok untuk bahasa Indonesia.

Modifikasi terdiri dari modifikasi di cluster aturan dan kondisi ukuran.Algoritma Porter hanya dapat melakukan akhiran pengupasan, beberapa tambahan harus dilakukan juga untuk menangani awalan pengupasan, *confix-stripping*, dan juga ejaan penyesuaian dalam kasus di mana pengenceran karakter pertama dari akar kata telah terjadi.

Stemming dalam bahasa Indonesia relative menantang. Ada variasi imbuhan termasuk *prefiks*, *sufiks*, *infiks*, dan *confixes*. Kata dalam bahasa Indonesia juga berasal dari perulangan kata, kombinasi imbuhan, dan kombinasi *afiks* dengan kata-kata diulang. Selain itu, bahasa Indonesia juga memiliki kata majemuk yang ditulis bersama-sama ketika melekat pada awalan dan akhiran. "*Confix-Stripping*" dimaksudkan untuk mengupas ke salah satu awalan atau akhiran. Selain berurusan dengan akhiran, beberapa pendekatan juga mencoba untuk menghapus prefiks umum yang tepat sebelum melakukan pengolahan bahasa lanjut.

Secara umum kesalahan dalam pemenggalan kata imbuhan sering kali terjadi karena kelalaian manusianya sendiri *(human error)*, kesalahan yang sering terjadi dalam pemenggalan kata imbuhan antara lain dalam hal pemakaian unit-unit kebahasaan kata, kalimat, paragraf, tanda baca dan ejaan serta kata-kata ambigu. Kesalahan kata juga dapat disebabkan oleh beberapa hal yang lain, seperti ketidaktahuan tentang penulisan. Kebanyakan kesalahan ini disebabkan oleh ketidaktahuan penulis mengenai kata atau penulisan yang benar, karena penulisan mereka hanya berdasar pada bunyi ejaan. Fadillah menyatakan penyebab-penyebab kesalahan pemenggalan kata di atas masih banyak penyebab kesalahan pemenggalan kata yang lain dalam hal penulisan atau pengetikan, Kata-kata yang muncul 80 % dalam dokumen-dokumen tidak berguna dalam proses retrival, penyisipan kata, penghilangan kata imbuhan dan perubahan awalan kata dasar yang mengalami perubahan bentuk serta partikel yang biasanya ditulis serangkai dengan kata dasarnya. Kata-kata ini disebut dengan istilah *stopword*s dan umumnya tidak dijadikan index term. Kandidat umum *stopword* adalah *article*, preposisi, dan konjungsi. Eliminasi *stopword*s bermanfaat dengan adanya pengurangan ukuran strukur index hingga 40%. Karena pengurangan ukuran index, beberapa kata kerja, kata sifat, dan kata keterangan lainnya dapat juga dapat dimasukkan juga ke dalam daftar *stopword*. Namun eliminasi *stopword*s dapat menyebabkan penurunan nilai *recall* (jumlah dokumen yang dihasilkan dan relevan/jumlah dokumen relevan) (Fadillah, 1999).

Salah satu penyesuaian dalam klasifikasi teks untuk suatu domain bahasa adalah dengan menyediakan suatu pendekatan *Confix-Stripping* yang spesifik pada bahasa tersebut.Sayangnya, perhatian terhadap domain Bahasa Indonesia masih tergolong minim, walaupun peringkat jumlah penduduknya terbanyak keempat di dunia dengan potensi pengguna Internet-nya yang semakin meningkat.*Confix-stripping* adalah salah satu diantara beberapa pendekatan *stemming* yangspesifik terhadap bahasa Indonesia.

METODE

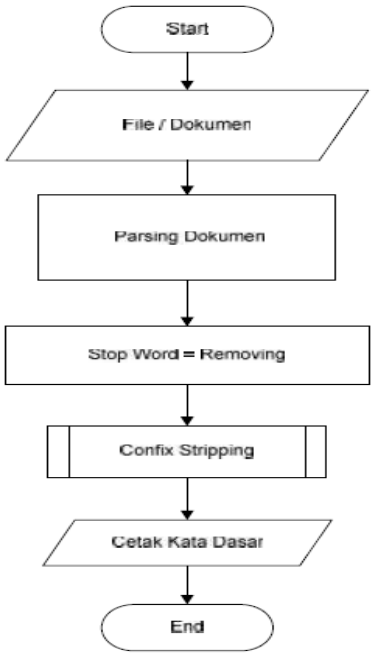
Dengan penerapan (algoritma) metode *Confix-Stripping*, maka proses pengambilan dan pengupasan kata imbuhan dokumen akan menjadi lebih cepatdan akurat dalam melakukan proses yang dilakukan.Proses *Stemming* merupakan pengelolaan *keyword* menjadi *keyword* yangutuh yaitu dengan menghilangkan imbuhan seperti diantaranya "yang", "di", "ke","me", "meng", "kan". Penguraian dari suatu kata menjadi bentuk kata dasarnya*(stem)*. Untuk lebih jelasnya tahapan proses *Stemming*

ISBN: 978-602-70604-1-8

ditunjukkan sebagai berikut:

- Token hasil *tokenizing* diperiksa apakah mengandung imbuhan atau tidak.
- Jika terdapat imbuhan maka akan dilakukan pembuangan imbuhan, terusberulang sampai tidak mengandung imbuhan.
- Jika tidak mengandung imbuhan maka akan ditampilkan.

Pada Gambar 1 menunjukkan tahapan yang dilakukan sistem dalam mencari kata dasar dari sebuah file dokumen yang diinputkan.



Gambar 1. *Flowchart* Sistem

Berdasarkan Gambar 1 dapat dijelaskan tahapan proses atau cara kerjasistem aplikasi untuk pencarian kata dasar teks berbahasa Indonesia denganpendekatan *Confix-Stripping* secara umum adalah sebagai berikut:

1) Proses 1 *(File/dokumen)*

Kata yang hendak di-*Stemming* dicari terlebih dahulu pada dokumen. Jika ditemukan, berarti kata tersebut adalah kata dasar, jika tidak maka proses 2 dilakukan.

2) Proses 2 *(Parsing/tokenizing)*

Tahapan ini akan melakukan pengecekan dari karakter pertama sampai dengan karakter terakhir. Apabila karakter ke (i) bukan merupakan pemenggal kata maka akan ditambahkan dengan karakter selanjutnya.Karakter pemenggal kata ini contohnya seperti tanda baca atau spasi.

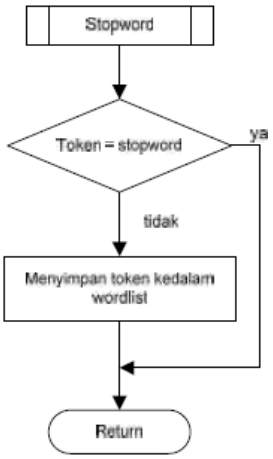
3) Proses 3 *(Stopword)*

Tahapan ini mengambil kata-kata penting dari hasil *token*. Bisa menggunakan algoritma *stop list* (membuang kata yang kurang penting) atau *word list* (menyimpan kata penting). Sistem ini menggunakan metode *stop list* yaitu penghilangan kata tidak penting *(stopword)* pada deskripsi melalui pengecekan kata-kata hasil *token* deskripsi apakah termasuk didalam daftar kata tidak penting *(stop list)* atau tidak. Jika termasuk didalam *stoplist* maka kata-kata tersebut akan di-*remove* dari

ISBN: 978-602-70604-1-8

deskripsisehingga kata-kata yang tersisa di dalam deskripsi di anggap sebagai katakatapenting atau *keywords (pattern)*. Tahapan proses *stopword* adalah sebagai berikut:

- Kata hasil token *Stemming* dibandingkan dengan tabel stopword.
- Dilakukan pengecekan apakah token sama dengan tabel stopwordatau tidak.
- Jika token sama dengan tabel stopword maka akan di-*remove*.
- Jika token tidak sama dengan tabel stopword akan ditampilkan. Yaitu menghasilkan token hasil *stopword* yang termasuk katapenting (*keyword*)

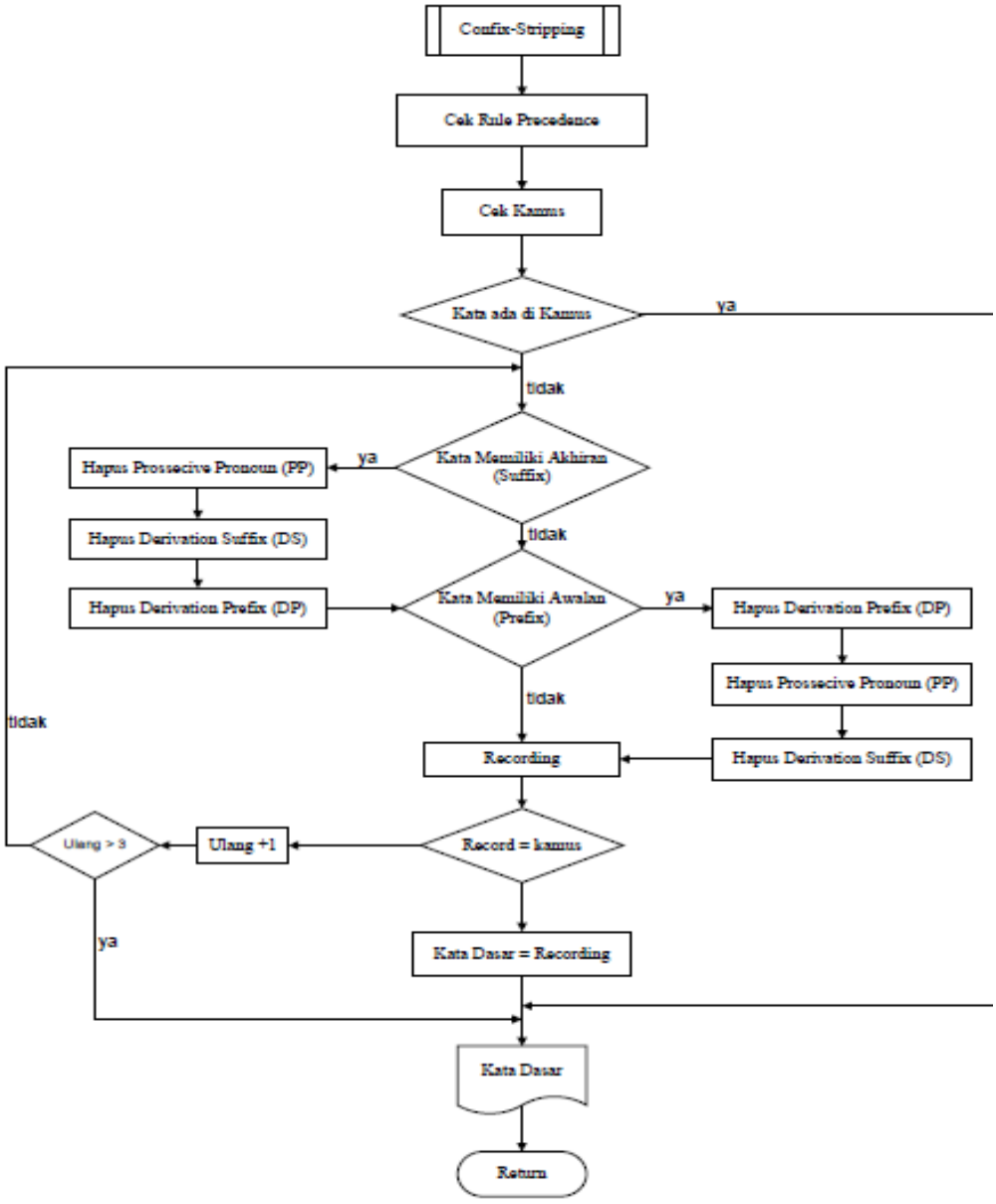


Gambar 2. *Flowchart Stopword*

4) Proses 4 *(Stemming Menggunkan Pendekatan Confix-Stripping)*

Pada Tahapan ini merupakan pengolahan *Stemming* dilakukan berdasarkaninput daftar filter term, proses *Stemming*ini menggunakan algoritma*Confix-Stripping*. Langkah pertama pada algoritma *stemmer* ini dilakukanpengecekan *rule Precedence* yakni larangan kombinasi awalan danakhiran, kemudian mencocokkan *term* pada elemen di*index* tertentudengan daftar "kata dasar" dalam *database* kamus. Jika cocok maka *term*tersebut langsung disimpan dalam variable *stemTerm*. Jika rule Precedencemengembalikan nilai benar proses pemenggalan akhiran dilakukan, jikatidak maka dilanjutkan dengan proses pemenggalan awalan. Kemudianproses recording yaitu proses penyesuaian kata dasar dengan aturanmengubah huruf pertama dari kata tersebut, apakah hasil recording samadengan kamus. Jika benar, maka kata dasar sama dengan hasil recordingjika tidak proses diulang. Proses perulangan ini dilakukan sampai batas 3x,jika proses berulang sampai pada batasan maka kata dasar sama denganhasil recording kemudian term yang ada langsung disimpan dalam variabeldan dianggap sebagai kata dasar. Alur *Stemming* dengan pendekatan *Confix-Stripping* dapat dilihat pada Gambar 3.

ISBN: 978-602-70604-1-8



Gambar 3. *Flowchart Stemming* dengan Pendekatan *Confix-Stripping*

HASIL DAN PEMBAHASAN

Pengujian yang digunakan untuk menguji sistem Stemming Artikel Berbahasa Indonesia ini adalah pendekatan *Confix-Stripping.* Pengujian adalahuntuk mengklasifikasi kebenaran dan kegagalan hasil *Stemming* pada *List box* aplikasi dari pengambilan kata secara otomatis pada Artikel Berbahasa Indonesia yang terdapat kata imbuhan kemudian akan diproses untuk dijadikan kata dasar.

Dimana pengujian-pengujian yang dilakukan adalah:

- Pengujian 1 pada *Stemming "di-kan" "di-nya" dan "di-kannya".*
  Dalam uji coba kesatu ini terdapat file dengan nama*coba 1.txt* yang berisikumpulan kata imbuhan *Confix-Stripping.* Uji coba menghasilkan data seperti pada Tabel 1.

ISBN: 978-602-70604-1-8

Tabel 1. Hasil Pengujian partikel *"di-kan", "di-nya",* dan *"di-kannya"*

No	Jenis Partikel	Input Kata	Kata Dasar	Hasil Stemming	Keterangan
1	Di-kan	Dibandingkan	Banding	Banding	Akurat
2		Dikumpulkan	Kumpul	Dikumpulkan	Tidak Akurat
3		Dikatakan	Kata	Kata	Akurat
4	Di-nya	Dikepalanya	Kepala	Kepalanya	Tidak Akurat
5		Ditangkapnya	Tangkap	Tangkapnya	Tidak Akurat
6		Dipipinya	Pipi	Pipinya	Tidak Akurat
7	Di-kannya	Ditemukannya	Temu	Temukannya	Tidak Akurat
8		Dibandingkannya	Bahagia	Bahagiakannya	Tidak Akurat
9		Dibayangkannya	Bayang	Bayangkannya	Tidak Akurat

- Pengujian 2 pada *Stemming "di-i" dan "diper-i"*
- Pengujian 3 pada *Stemming "Meng-kan", "Mem-kan", "Memper-kan","Meny-i", "Meny-kan", "Men-kan", "Me-kan", dan "Meng-nya"*
- Pengujian 4 pada *Stemming "Memper-i", "Men-i", dan "Meng-i"*
- Pengujian 5 pada *Stemming "Ber-nya" dan "Ber-an"*
- Pengujian 6 pada *Stemming "ke-an" dan "ke-annya"*
- Pengujian 7 pada *Stemming "peng-an" "peny-an" "per-an" "pen-an"dan "pe-an"*
- Pengujian 8 pada *Stemming "Ber-", "Di-", "Men-", "Meng-", "Meny-"dan "Mem-"*
- Pengujian 9 pada *Stemming "Pen-", "Pem-", "Peng-", dan "Ter-"*
- Pengujian 10 pada *Stemming "-nya", "-lah", "-pun", "-kah", "-mu", "-an", "-kan", dan "-i"*

Setelah melakukan 10 pengujian kebenaran dan kegagalan pencarian kata dasarpada beberapa contoh gabungan awalan-akhiran (*Confix-Strippng),* awalan*(Prefix),* dan akhiran *(Suffix),* maka dapat dihitung nilai akurasi dan *running time* hasil *Stemming.*

Tabel 2. Pengujian Akurasi Hasil *Stemming*

No	Nama file	Size file dokumen (kb)	Keterangan jumlah kata		% Akurasi	Waktu (detik)
			File Dokumen	Hasil yang tidak akurat		
1	Test1.docx	235	100	4	94	13
2	Test2.docx	28	100	7	93	14
3	Test3.docx	70	100	5	95	13
4	Test4.docx	28	100	4	96	13
5	Test5.docx	168	100	6	94	25
6	Test6.docx	28	100	6	94	16
7	Test7.docx	1	100	6	94	25
8	Test8.docx	172	100	4	96	17
9	Test9.docx	28	100	4	96	13
10	Test10.docx	1	100	6	94	23
11	Test11.docx	12	125	5	96	15
12	Test12.docx	13	150	7	95	18
13	Test13.docx	446	175	8	95	22

ISBN: 978-602-70604-1-8

No	Nama file	Size file	Keterangan jumlah kata		%	Waktu
14	Test14.docx	19	200	10	95	22
15	Test15.docx	13	225	15	93	33
16	Test16.docx	21	680	32	95	2mt 21dt
17	Test17.docx	13	250	7	97	39
18	Test18.docx	14	275	17	93	41
19	Test19.docx	15	300	16	94	38
20	Test20.docx	16	325	13	96	1mt 7dt
RATA-RATA AKURASI YANG DIDAPAT					**94,85%**	

KESIMPULAN DAN SARAN

Setelah melakukan penelitian dengan melakukan aktifitas perancangan, pembuatan Aplikasi dan pengujian, dapat ditarik kesimpulan bahwa:

1. Pencarian kata dasar secara otomatis pada dokumen text berbahasa Indonesia dapat mempercepat dan memudahkan untuk proses pencarian kata dasar dari dokumen teks Berbahasa Indonesia.
2. Running Time Stemming pencarian kata dasar kedalam sebuah dokumen text berbahasa Indonesia tidak dipengaruhi oleh besar kecilnya ukuran (*size*) file pada input dokumen, tetapi dipengaruhi oleh banyaknya jumlah kata dalam dokumen tersebut.
3. Hasil pengujian yang dilakukan terhadap 20 dokumen teks berbahasa Indonesia yang diuji didapatkan rata-rata nilai akurasi sebesar 94.8%.

Adanya kekuranga-kekurangan dalam penelitian yang telah dilakukan ini, sehingga bisa dikembangkan dan disempurnakan lagi, antar lain:

1. Dalam penelitian ini hanya bisa mengolah dokumen dalam format .doc, dan .rtf. Sehingga penelitian dapat dikembangkan untuk inputan dokumen dengan format lain (.pdf, .xls, dan lain-lain)
2. Penelitian dapat dikembangkan dengan penyempurnakan proses stemming untuk kata-kata yang mengandung partikel “dikan”, “ke-nya”, “di-nya”, “memper-kan”, “memper-i”,“berpeng-an”, “memper-kan”, “ke-annya”, “ke-nya”, “peng-an”, “pen-an”, “peny-kan”, “peny-an”, “per-kan”, “perkan”, dan kata imbuhan sisipan.

DAFTAR PUSTAKA

Abdelmalek A. (2007). *Evaluation and Comparison of Concept Based and N-Grams Based Text Clustering using SOM*.TIMC-IMAG Laboratory IN3S, Joseph Fourier University.

Arifin, A. Z. dan A. N. Setiono. (2002). Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma *Single Pass Clustering*.*Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA)*, Teknik Elektro, Institut Teknologi Sepuluh Nopember.

Asian J. (2007). *Effective Techniques for Indonesian Text Retrieval*. PhD Thesis School of Computer Science and Information Technology RMIT University Australia.

ISBN: 978-602-70604-1-8

B. A. A. Nazief and M. Adriani, (1996). *Confix-stripping: Approach to stemming algorithm for* Bahasa Indonesia. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta.

C. D. Paice, (1994). *An evaluation method for stemming algorithms*. In *Proceedings* of the ACM- SIGIR International Conference on Research and Development in Information Retrieval, pages 42–50, Dublin, Ireland. Springer-Verlag New York, Inc.

Elizabet N. S. C. P. (2013). Rancangan Bangunan Aplikasi ChatBot Informasi Objek Wisata Kota Bandung dengan pendekatan *Natural Language Processing*. Universitas Komputer Indonesia, Bandung.

Fadillah Z Tala.(1999). *A Study of Stemming Effect on Information Retrieval in* Bahasa Indonesia".

G. Salton.(1968). *Computer evaluation of indexing and text processing.Journal of the ACM*.

H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. Moeliono. (1998). Tata Bahasa Baku Bahasa Indonesia. Balai Pustaka, Jakarta, Indonesia, Third edition.

I Putu Adhi Kerta Mahendra, Agus Zainal Arifin, Henning Titi Ciptaningtyas. (2008). *Penggunaan Algoritma Semut dan Confix-Stripping Stemmer untuk Klasifikasi Dokumen teks Berbahasa Indonesia*. Jurnal

James Suciadi. (2001). Studi Analisis Metode-Metode Parsing dan Interpretasi Semantik Pada Natural Language Processing. JURNAL INFORMATIKAVol. 2, No. 1:13 – 22.

Kridalaksana, Harimurti. (1996). *Pembentukan Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama.

M. Popovi˘c and P. Willett. (1992). The effectiveness of stemming for natural language access to Slovene textual data. Journal of the American Society for Information Science, 43(5): 384–390.

N. Idris. (2001). Automated essay grading system using nearest neighbor technique in information retrieval. Master's thesis, University of Malaya, Malaysia.

R. Krovetz. (1993) Viewing Morphology as an Inference Process. In *Proceedings of the16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–203, Pittsburgh, Pennsylvania.ACM Press.

Setiono, Ari Novan. (2001). Implementasi Aplikasi Information Retrieval untuk Pendeteksian dan Klasifikasi Berita Kejadian Berbahasa Indonesia Berbasis Web. Tugas Akhir, Teknik Informatika, Institut Teknologi Sepuluh Nopember Surabaya.

W. B. Frakes and R. Baez. (1992). *Information Retrieval, Data Structures and Algorithms*. Prentice Hall.

Xu, J. and Croft, W. B. (1998). *Corpus-based stemming using cooccurrence of word variants*. ACM Transactions on Information Systems, Vol 16 No.1:61-81.

ISBN: 978-602-70604-1-8