Jurnal Humaniora dan Teknologi
Volume 10, Nomor 2, Oktober 2024
p–ISSN: 2443–1842
e–ISSN: 2614–3682

PENERAPAN ALGORITMA IDRIS PADA DOKUMEN DENGAN MENGGUNAKAN TEKS BAHASA INDONESIA

Nindy Permatasari[1], Cahya Karima[2]
Politeknik Negeri Tanah Laut
E–mail: nindy@politala.ac.id[1], cahyakarima@politala.ac.id[2]

Abstrak

Meningkatnya jumlah pencarian informasi pada saat ini sangat relevan juga dengan banyaknya dokumen yang ada, tentu saja hal ini menyebabkan pengguna akan kesulitan untuk dapat menemukan dokumen yang relevan sesuai dengan query yang diinputkan oleh pengguna. Metode stemming adalah salah satu metode yang dapat mengatasi masalah tentang pencarian informasi di dalam dokumen secara relevan dan dapat digunakan untuk mengurangi perbedaan dari bentuk suatu kata dengan cara mengembalikan kata tersebut ke dalam bentuk dasar. Algoritma Idris merupakan salah satu algoritma untuk metode stemming yang merupakan pengembangan dari algoritma Ahmad, Yusoff dan Sembok. Penelitian ini bertujuan untuk dapat menganalisis tingkat keberhasilan dari Algoritma Idris dalam melakukan stemming pada dokumen teks Bahasa Indonesia dikarenakan algoritma ini dibangun untuk stemming pada Bahasa Melayu. Kemudian penelitian ini dilakukan dengan menganilisis penerapan Algoritma Idris pada dokumen Bahasa Indonesia. Setelah dilakukan pengujian dengan menggunakan sepuluh dokumen pada Algoritma Idris persentase keberhasilan steming dengan menggunakan algoritma idris adalah 83% dengan ICF 0,1628 dan waktu untuk pemrosesan algoritma untuk melakukan stemming adalah 35,865 detik.

Kata kunci : Algoritma Idris, Bahasa Indonesia, Stemming

Abstract

The increasing number of search information at this time is also very relevant to the number of existing documents, of course this causes users to have difficulty finding relevant documents according to the query entered by the user. The stemming method is one method that can solve the problem of searching for information in documents relevantly and can be used to reduce the differences in the form of a word by returning the word to its basic form. The Idris algorithm is one of the algorithms for the stemming method which is a development of the Ahmad, Yusoff and Sembok algorithms. This study aims to be able to analyze the success rate of the Idris Algorithm in stemming Indonesian text documents because this algorithm was built for stemming in Malay. Then this study was conducted by analyzing the application of the Idris Algorithm to Indonesian documents. After testing using ten documents on the Idris Algorithm, the percentage of stemming success using the Idris algorithm was 83% with an ICF of 0.1628 and the time to process the algorithm to perform stemming was 35.865 seconds.

Keywords : Idris Algorithm, Indonesian, Stemming

PENDAHULUAN

Suatu proses untuk dapat menemukan sebuah kata dasar dari sebuah kata yang dilakukan dengan cara menghilangkan semua imbuhan (afiks) baik itu yang terdiri dari awalan (prefix), sisipan (infiks), akhiran (suffiks) dan kombinasi awalan-akhiran (konfiks) yang terdapat pada kata turunan merupakan suatu proses yang disebut dengan stemming. Stemming merupakan inti dari sebuah Teknik untuk melakukan pemrosesan *natural language* sehingga mendapatkan kembali informasi atau disebut juga dengan *information retrieval* (Noverdy, 2015). Sedangkan *information retrieval* merupakan system yang mana user akan dapat memasukkan suatu query tertentu kemudian system dapat mengembalikan nilai dari suatu informasi tersebut dapat berupa dokumen maupun data yang memiliki relevansi terhadap query

p–ISSN: 2443–1842
e–ISSN: 2614–3682

yang telah diberikan oleh user tersebut.

Saat ini kebutuhan pengguna akan pencarian informasi semakin meningkat dan jumlah dokumen teks yang dapat diakses juga semakin banyak, hal ini dapat mengakibatkan pengguna semakin sulit menemukan dokumen yang relevan dengan query yang dimasukkan (Utomo, 2013). Indikator yang biasanya digunakan dalam menentukan nilai relevansi dari suatu pencarian pada sebuah dokumen adalah bagaimana kesesuaian antara query yang diberikan dengan dokumen yang dihasilkan pada pencarian. Misalkan pada kasus berikut ini dengan *stemming* bahasa Indonesia, mencari suatu dokumen dengan menggunakan judul "baca buku" dan menggunakan sebuah *query* "membaca", maka dokumen yang tidak akan mendapatkan hasil yang sesuai dengan pencarian. Namun, dengan kita menggunakan sebuah metode *stemming*, maka *query* seperti itu "membaca" serta "dibaca" akan dianggap mempunyai interpretasi yang sama yaitu sebuah "baca" sehingga antara kata pada dokumen dengan *query* bisa cocok. Dengan menggunakan metode diatas maka pencarian dokumen akan berhasil (Tala, 2003). S*temming* dengan menggunakan Bahasa Indonesia terdapat beberapa teknik yang dapat digunakan seperti Jelita Asian tahun 2005, Arifin & Setiono tahun 2002, Nazief & Adriani tahun 1996, Ahmad Yusoff Sembok tahun 1996, Vega tahun 2001, Idris tahun 2001 dan *ECS Stemmer* tahun 2008. Teknik *stemming* tersebut dikembangkan untuk sebuah alasan agar dapat mereduksi sebuah *term* menjadi sebuah bentuk dasarnya saja (Tuhpatussania, Ema, & Anggit, 2022) (Wardana, Iswara, & Banu, 2019) (Wahyudi, Teguh, & Didik, 2017) (Magriyanti, 2018).

Metode stemming digunakan untuk mengatasi masalah pencarian informasi yang tersimpan didalam dokumen secara efektif dan efisien dan juga digunakan untuk mengurangi perbedaan bentuk dari suatu kata dengan mengembalikannya ke dalam bentuk kata dasar. Teknik stemming dikembangkan untuk alasan mereduksi term menjadi bentuk dasarnya. Dalam penelitian ini akan dilakukan analisis algoritma dalam metode Stemming yaitu Algoritma Idris yang dapat diterapkan untuk pencarian dokumen teks bahasa Indonesia.Oleh karena itu, maka penelitian terhadap analisis performansi dari algoritma ini dilakukan dengan tujuan mendapatkan informasi dari algoritma yaitu berupa informasi kecepatan dan akurasi serta jumlah langkah dari Algoritma Idris dalam penerapannya pada sebuah simulator.

METODE PENELITIAN

A. *Stemming*

Stemming merupakan suatu proses pembuangan atau pemotongan afiks (baik prefiks maupun sufiks) untuk mendapatkan bentuk akar atau dasar dari suatu *term* (berupa kata). Dalam konteks *information retrieval*, *stemming* digunakan untuk mereduksi bentuk *term* untuk menghindari ketidakcocokan. Hal ini dikarenakan afiks dapat mengandung informasi seperti bagian dari percakapan, *plurarity*, dan sebagainya. Perlu diingat bahwa proses *stemming* di sini bukan merupakan pekerjaan dalam *etimologi grammar*, sehingga dapat ditoleransi apabila algoritma *stemming* menghasilkan kata yang tidak bermakna. Algoritma *stemming* dapat dibedakan menjadi *context-free* dan *context-sensitive*. *Context-free* membuang akhiran tanpa adanya suatu batasan, sedangkan *context-sensitive* melibatkan banyak batasan kontekstual untuk mencegah pembuangan akhiran yang mengakibatkan *stem* yang dihasilkan menjadi rusak.

p–ISSN: 2443–1842
e–ISSN: 2614–3682

B. *Stemming* Bahasa Indonesia

Morfologi adalah bagian dari ilmu bahasa yang membicarakan atau mempelajari seluk-beluk bentuk kata serta pengaruh perubahan bentuk kata terhadap golongan dan arti kata. Kata yang dibentuk dari kata lain pada umumnya mengalami tambahan bentuk pada kata dasarnya. Perubahan-perubahan bentuk kata menyebabkan adanya perubahan golongan dan arti kata. Tiga macam proses morfologis, yaitu pertama, bergabungnya morfem bebas dengan morfem terikat disebut afiksasi. Kedua, pengulangan morfem bebas disebut reduplikasi, dan ketiga, bergabungnya morfem bebas dengan morfem bebas disebut pemajemukan. Pada proses yang pertama menghasilkan kata berimbuhan, yang kedua menghasilkan kata ulang, dan yang ketiga menghasilkan kata majemuk. Imbuhan (afiks) adalah bentuk (morfem) terikat yang dipakai untuk menurunkan kata.Pada umumnya imbuhan (afiks) hanya dikenal ada empat, yaitu awalan (prefiks), sisipan (infiks), akhiran (sufiks), awalan dan akhiran (konfiks).

C. Algoritma Idris

Algoritma Idris merupakan pengembangan dari algoritma sebelumnya, yaitu algoritma Ahmad, Yusoff, dan Sembok yang awalnya dikembangkan untuk Bahasa Melayu[9]. Algoritma ini memperluas skema algoritma Ahmad, Yusuf dan Sembok untuk *stemming* dan *recoding* dalam dokumen teks Bahasa Melayu. Kemiripannya dengan dokumen teks Bahasa Indonesia, membuat algoritma Idris dapat diterapkan juga untuk dokumen teks berbahasa Indonesia. Algoritma ini menghapus prefiks dan sufiks sampai akar kata ditemukan dalam kamus. Dalam tugas akhir ini, algoritma Idris mengadopsi asumsi Arifin dan Setiono bahwa kata-kata bahasa Indonesia dapat memiliki paling banyak dua awalan (prefiks) dan tiga akhiran (sufiks). Aturan-aturan (*rules*) dalam algoritma Idris, diadopsi dari *rules* pada algoritma Othman, yaitu:

(1) Prefiks rules format : Prefiks+
Contoh : di + jajah → dijajah
(2) Suffiks rules format : +Suffiks
Contoh : jajah + an → jajahan
(3) Infix rules format : +Infix+
Contoh : tapak (+el+) → telapak
(4) Prefiks-Suffiks pair rules format : Prefiks+Suffiks
Contoh : meN + takluk → menakluki

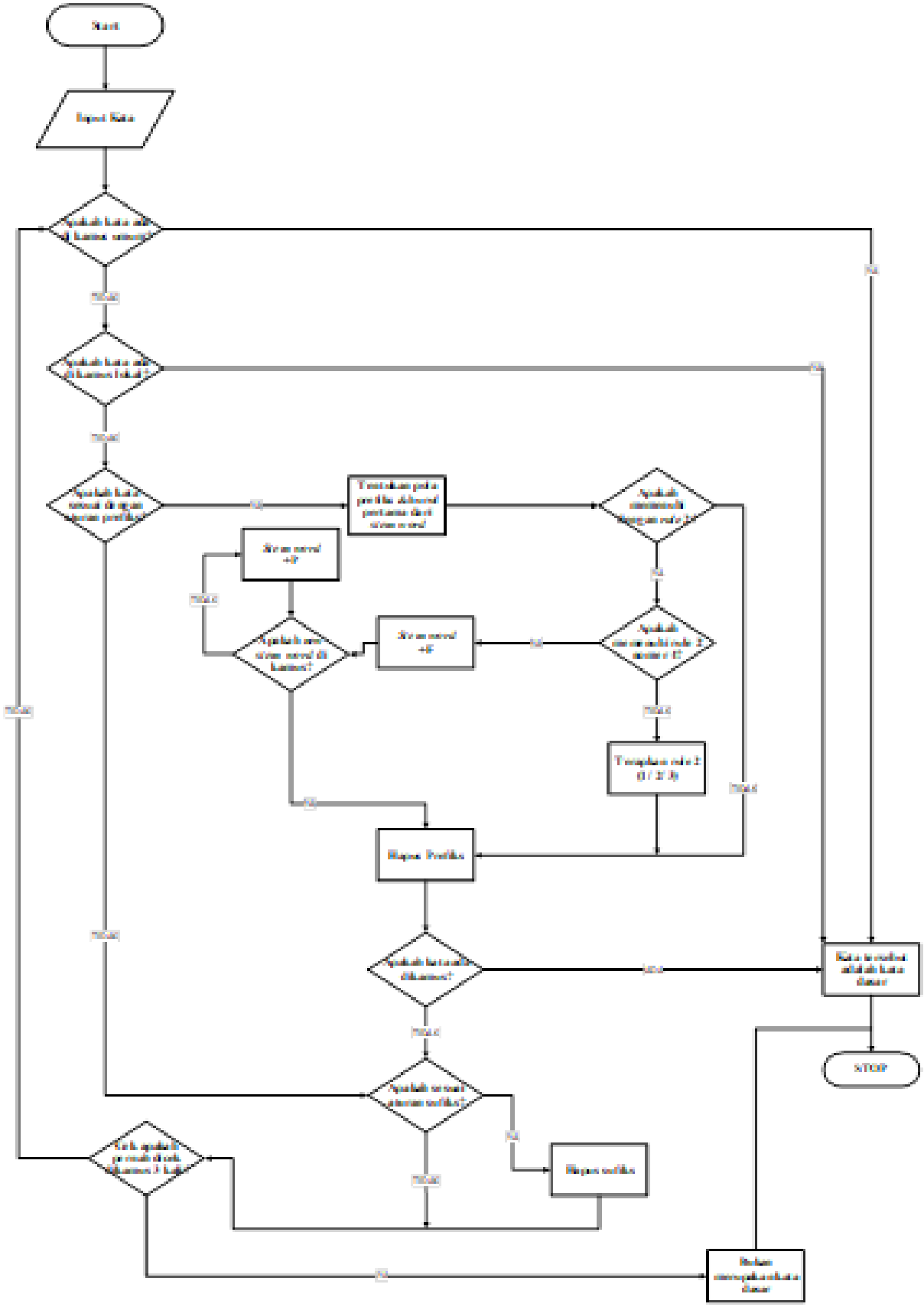
Jika dalam kenyataannya, hasil *stemming* dari *rules* di atas tidak lengkap, maka cek huruf pertama dari kata tersebut. Aturan ini biasa disebut Rule 2. Jika huruf pertama adalah huruf hidup (vokal), maka :

(1) Tambahkan *t* setelah menghilangkan imbuhan *men-* atau *pen-*
(2) Tambahkan *k* setelah menghilangkan imbuhan *meng-* atau *peng-*
(3) Tambahkan *s* setelah menghilangkan imbuhan *meny-* atau *peny-*
(4) Tambahkan *f* atau *p* setelah menghilangkan imbuhan *mem-* atau *pem-*

Karena didesain untuk Bahasa Melayu, algoritma ini menggunakan prefiks dan sufiks yang sedikit berbeda dari Bahasa Indonesia, yang dapat berpeluang menjadi *overstemming*. Secara garis besar, skema algoritma Idris dapat dilihat pada Gambar 1

p–ISSN: 2443–1842
e–ISSN: 2614–3682

berikut ini



Gambar 1. Skema Algoritma Idris

p–ISSN: 2443–1842
e–ISSN: 2614–3682

Berdasarkan skema pada Gambar 1. dapat dijelaskan bahwa skema dari Algoritma Idris adalah :

(1) Cek kata yang dicari dalam kamus umum. Jika kata tersebut ada dalam kamus umum,maka kata tersebut merupakan kata akar (kata dasar) dan keluar dari algoritma. Jika tidak ada, maka dilanjutkan ke langkah (2).
(2) Cek kata yang dicari dalam kamus khusus. Jika kata tersebut ada dalam kamus khusus, maka kata tersebut merupakan kata akar (kata dasar) dan keluar dari algoritma. Jika tidak ada, maka dilanjutkan ke langkah (3).
(3) Cek kata pada aturan *prefiks*. Jika kata tersebut cocok dengan aturan *prefiks*, cek pola dari *prefiks* tersebut dan huruf pertama dari kata yang akan di-*stem*. Jika tidak cocok, makan dilanjutkan ke langkah (4).
(4) Jika pola *prefiks* pada kata cocok dengan *Rules* 2, maka ubah kata tersebut sesuai dengan *rule*. Jika tidak cocok, maka hilangkan *prefiks* dan lanjutkan ke langkah (7).
(5) Cek *prefiks* dari kata tersebut, cocokkan polanya pada *rules* 2. Jika cocok dengan keempat *rule* tersebut, lalu cek kata yang telah di-stem pada kamus dan lanjutkan ke langkah (6). Jika tidak cocok, maka hilangkan prefiks dan lanjutkan ke langkah (7).
(6) Cek prefiks dari kata tersebut, cocokkan polanya pada *rules* 2. Jika cocok dengan keempat *rule* tersebut, lalu cek kata yang telah di-*stem* pada kamus dan lanjutkan ke langkah (6). Jika tidak cocok, maka hilangkan prefiks dan lanjutkan ke langkah (7). Jika kata tersebut tidak ada pada kamus, maka kembali ke langkah (5). Jika cocok, maka hilangkan prefiks dan lanjutkan ke langkah (7).
(7) Cek kata telah cocok pada kamus, lalu kata tersebut merupakan kata dasar, lalu keluar dari algoritma. Jika tidak cocok, maka lanjutkan ke langkah (8).
(8) Cek kata pada aturan suffiks. Jika cocok dengan aturan tersebut, maka hilangkan suffiks dan lanjutkan ke langkah (1). Jika tidak cocok, lanjutkan juga ke langkah (1).

Rule–rule yang digunakan pada algortima Idris dapat dilihat pada Tabel 1. Berikut ini:

Tabel 1. Tabel *Rule* Algoritma Idris

Kelas Affiks	*Affixes*	Contoh kata
Prefiks	di-	dirantau
	ke-	kemana
	ber-	berlari
	men-	mencari
	ter-	terdapat
	pen-	pendapat
	per-	pertanda
Suffiks	-an	jalanan
	-i	jalani
	-kan	jalankan
	-nya	punyanya
	-lah	apalah
	-kah	inikah
Prefiks - Suffiks	ber – an	bertaburan
	ber – kan	berdasarkan

p–ISSN: 2443–1842
e–ISSN: 2614–3682

Kelas Affiks	*Affixes*	Contoh kata
	di – i	dijauhi
	ke – an	kedatangan
	men – kan	mendengarkan
	men – i	mendapati
	memper – i	mempertanyai
	memper – kan	memperdengarkan
	per – an	perdagangan
	se – nya	seharusnya
Infiks	el	telapak
	em	gelembung
	er	serabut
	in	kinerja

PENGUJIAN

Tahapan pengujian yang akan dilakukan terdiri dari pengujian dari Algoritma Idris. Pengujian ini dilakukan dengan tujuan untuk mengetahui performansi dari algoritma Idris digunakan untuk menstemming kata baku dalam bahasa Indonesia yang akan dilakukan pada simulator yang telah dibuat, pengujian meliputi akurasi dan waktu proses yang akan dilakukan untuk menghasilkan *output*. Skenario pengujian sangat diperlukan agar proses pengujian yang dilakukan dapat mencapai tujuan yang diinginkan. Adapun skenario pengujian yang dilakukan yaitu melakukan proses pencarian kata dasar pada *database* dan menganalisis hasil *stemming*. Berikut deskripsi mengenai skenario pengujian yang dilakukan adalah sebagai berikut.

a. Skenario Pengujian 1

Menguji *stemmer strength* dari Algoritma Idris. Pengujian dilakukan dengan menganalisis nilai yang dihasilkan berupa nilai jumlah kata sebelum dan sesudah *stemming* dengan nilai icf dan persentase. Skenario yang dilakukan yaitu dengan menggunakan 10 dokumen [5] dan menganalisis nilai parameter yang dihasilkan. Berikut adalah skenario pengujian pertama yang dilakukan.

1. Menggunakan kata-kata sesuai dengan dokumen yang telah ditentukan. Pada pengujian skenario 1 ini.
2. Menghasilkan *output* berupa kata dasar setelah proses *stemming*.
3. Menghitung nilai parameter icf dan persentase dari masing-masing dokumen.

b. Skenario Pengujian 2

Menguji keakuratan Algoritma Idris terhadap stem yang dihasilkan. Skenario yang dilakukan, yakni menggunakan 10 dokumen yang telah ditentukan sebelumnya dan menganalisis isi dokumen uji berdasarkan jumlah kata yang tidak berhasil distemming. Berikut adalah skenario pengujian kedua yang dilakukan.

1. Menggunakan kata-kata sesuai dokumen yang telah ditentukan.
2. Menghasilkan *output* berupa kata-kata setelah dilakukan proses *stemming*.
3. Menganalisis dan menghitung jumlah kata yang tidak berhasil distemming dengan benar.

p–ISSN: 2443–1842
e–ISSN: 2614–3682

HASIL DAN PEMBAHASAN

A. Skenario Pengujian 1

Berikut merupakan tabel hasil dari pengujian kesepuluh dokumen di atas, dapat dilihat pada Tabel 2 berikut ini.

Tabel 2. Hasil pengujian dokumen *stemming* dengan Idris

No	Nama	Sebelum	Algoritma Idris			
			Setelah	ICF	s	%
1	Dok 1	239	209	0,12552	4,37	87%
2	Dok 2	261	227	0,13027	4,51	86%
3	Dok 3	323	276	0,14551	6,06	85%
4	Dok 4	511	415	0,18787	8,75	81%
5	Dok 5	608	512	0,15789	12,87	84%
6	Dok 6	1023	852	0,16716	20,93	83%
7	Dok 7	2224	1850	0,16817	38,3	83%
8	Dok 8	3653	3109	0,14892	64,49	85%
9	Dok 9	4464	3746	0,16094	92,37	84%
10	Dok 10	5815	4450	0,23474	106	76%
Rata - rata				0,1627	35,9444	83%

B. Skenario Pengujian 2

Pada pengujian ke skenario 2 ini dilakukan untuk mengetahui keakuratan hasil *stemming* dari algoritma Idris. Kata–kata yang gagal *stemming* dapat disebabkan oleh skema pemotongan imbuhan yang tidak dapat mencari kata dasar walaupun imbukan sudah dipotong dan sudah dikembalikan jika mengalami salah imbuhan. Berikut ini merupakan contoh kata yang tidak berhasil distemming oleh Algoritma Idris dapat dilihat pada Tabel 3 di bawah ini.

Tabel 3. Contoh kata yang tidak berhasil distemming

No	Kata	Hasil Stemming	Kata dasar yang benar
1	geletar	geletar	getar
2	mengantuk	antuk	kantuk
3	serabut	serabut	sabut
4	temali	temali	tali
5	kinerja	kinerja	kerja
6	mengantuk	antuk	kantuk
7	mengarang	arang	karang

Kemudian setelah dilakukan analisis kembali pada kata–kata yang telah berhasil distemming tetapi terdapat beberapa kesalahan yang ditemukan karena kata dasar tersebut mengalami *overstemming* atau pemotongan huruf yang seharusnya tidak dilakukan atau kelebihan memotong huruf sehingga merubah kata dasar yang seharusnya. Berikut contoh pemenggalan imbuhan yang berlebihan atau yang seharusnya tidak dilakukan pemotongan dapat dilihat pada Tabel 4 di bawah ini.

p–ISSN: 2443–1842
e–ISSN: 2614–3682

Tabel 4. Kata yang mengalami kesalahan *stemming*

No	Kata	Hasil Stemming	Kata dasar yang benar	*Tipe Kesalahan*
1	negeri	neger	negeri	*overstemming*
2	buat	bu	buat	kesalahan pemotongan
3	kelas	as	kelas	*overstemming*
4	sampai	sampa	sampai	*overstemming*
5	mengenali	nali	kenal	*overstemming*

Dalam melakukan *stemming* dengan algoritma Idris ada beberapa kata yang memang tidak bisa distemming hal ini dikarena kata tersebut merupakan kata dasar yang belum tercantum pada saat pembuatan kamus kata dasar atau kata tersebut bukan kata dasar yang tercantum dalam kamus bahasa Indonesia. Kata – kata seperti diatas menyebabkan *term* tidak dapat distemming dengan benar. Berikut merupakan kata yang tidak terdaftar dalam kamus dapat dilihat pada Tabel 5 di bawah ini.

Tabel 5. Kata yang tidak terdaftar dalam kamus

Kata Dasar		
markup	hembus	bongkok
kastil	astronot	akupuntur
negeri	lesat	bolpen
akting	klon	congkel
aksesoris	fesyen	
masing	nampak	

KESIMPULAN

Seperti yang dapat dilihat dari Tabel 3 dan Tabel 4 maka dapat disimpulkan bahwa semakin sedikit kata yang berhasil distemming maka nilai icf akan semakin besar. Nilai icf algoritma Idris artinya bahwa algoritma tersebut dapat diterapkan dengan baik untuk *stem* bahasa Indonesia Berdasarkan waktu pemrosesan algoritma Idris dalam menstemming rata–atanya 35,94 detik. Jadi dapat ditarik kesimpulan bahwa berdasarkan akurasi dan kecepatan algoritma Idris baik digunakan untuk menstemming dokumen teks bahasa Indonesia. Akan tetapi algoritma Idris memiliki kekurangan dalam hal gabungan kata yang merupakan gabungan dari dua kata yang menjadi satu seperti kata “prasejarah” merupakan kata yang berasal dari “pra” dan “sejarah”. Kata seperti ini tidak dapat diatasi dengan menggunakan aturan pemotongan imbuhan tetapi perlu dilakukan penyesuai algoritma baru untuk dapat menyelesaikan permasalahan gabungan dua kata yang menjadi satu.

DAFTAR PUSTAKA

A, A. s., Yudi, P., & Arvita, E. (2019). Stemming Analysis Indonesian Language News Text with Porter Algorithm. *Journal of Physics: Conference Series*.

Ahmad, F., Mohammed, Y., & Tengku, M. T. (1996). Experiments with a stemming algorithm for Malay words. *Journal of the American Society for Information Science*, 909-918.

p–ISSN: 2443–1842
e–ISSN: 2614–3682

Idris, N., & Mustapha, S. S. (n.d.). STEMMING FOR TERM CONFLATION IN MALAY TEXTS. *International Conference on Artificial Intelligence (IC-AI 2001).*

Magriyanti, A. A. (2018). ANALISIS PENGEMBANGAN ALGORITMA PORTER STEMMING. *https://doi.org/10.31227/osf.io/7ge4v*.

Noverdy, A. (2015). *Implementasi Modifikasi Algoritma Enhanced Confix Stripping Stemmer pada Teks Bahasa Indonesia.* Bandung: Tel-U Collection.

Rifai, W., & Winarko, E. (2019). Modification of Stemming Algorithm Using A Non Deterministic Approach To Indonesian Text. *Indonesian Journal of Computing and Cybernetics System*.

Tala, F. Z. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Amsterdam: Institute for Logic, Language and Computation Universiteit van Amsterdam.

Tuhpatussania, S., Ema, U., & Anggit, D. H. (2022). COMPARISON OF PORTER'S STEMMING ALGORITHM AND NAZIEF & ADRIANI'S STEMMING ALGORITHM IN DETERMINING INDONESIAN LANGUAGE LEARNING MODULES. *PILAR Nusa Mandiri : Journal of Computing and Information System*, 203-210.

Utomo, M. S. (2013). Implementasi Stemmer Tala pada Aplikasi Berbasis Web . *Jurnal Teknologi Informasi DINAMIK*, 41-45.

Wahyudi, D., Teguh, S., & Didik, N. (2017). IMPLEMENTASI DAN ANALISIS ALGORITMA STEMMING NAZIEF & ADRIANI DAN PORTER PADA DOKUMEN BERBAHASA INDONESIA. *Jurnal Ilmiah SINUS*, 49-56.

Wardana, H. K., Iswara, S., & Banu, W. Y. (2019). Sistem Pemeriksa Pola Kalimat Bahasa Indonesia berbasis. *JNTETI*, 211-217.