ANALISIS DAN PERBANDINGAN STEMMING ALGORITMA PORTER DENGAN ALGORITMA AHMAD YUSOFF SEMBOK DALAM DOKUMEN TEKS BAHASA INDONESIA

Arif Siswandi dan Nurhadi Surojudin
Universitas Pelita Bangsa
Jl. Inspeksi Kalimalang No.9, Cibatu, Cikarang Pusat, Bekasi, Jawa Barat 17530 {arif.siswandi, nurhadi}@pelitabangsa.ac.id

ABSTRAK

Stemming adalah proses untuk mengklasifikasi berbagai macam variasi morfologikal dari sebuah kata maupun kalimat menjadi satu bentuk dasar yang sama. Di dalam stemming berbahasa Indonesia, terdapat dua jenis metode stemming yang sudah ada, yaitu algoritma stemming yang berbasis kamus (dictionary based) dan algoritma stemming yang berbasis non-kamus (purely rule based). Penelitian ini menggunakan model perbandingan dua algoritma stemming berbasis kamus dan algoritma stemming menggunakan aturan imbuhan. Algoritma yang digunakan adalah algoritma Porter Indonesia untuk yang berbasis kamus. Algoritma stemming berbasis aturan imbuhan yang digunakan adalah algoritma Ahmad Yusoff Sembok. Pengujian dilakukan dengan menggunakan 100 dokumen teks Bahasa Indonesia yang sudah ditentukan sebelumnya. Hasil pengujian yang dilakukan menunjukkan bahwa nilai Akurasi yang paling tinggi terdapat pada algoritma Porter, nilai Overstemming dan Understemming yang paling sedikit juga terdapat pada algoritma Porter. Dari pengujian yang telah dilakukan menunjukkan bahwa algoritma Porter lebih baik daripada Algoritma Ahmad Yusoff Sembok..

Kata Kunci :*Information Retrieval, Stemming, Akurasi, Overstemming dan Understemming*

PENDAHULUAN

Dalam melakukan pencarian informasi berupa dokumen berbentuk teks atau yang dikenal dengan istilah *Information Retrieval* (IR) adalah merupakan proses pemisahan dokumen-dokumen yang dianggap relevan dari sekumpulan dokumen yang tersedia. Dengan penambahan jumlah dokumen teks yang dapat diakses di internet yang diikuti dengan peningkatan kebutuhan pengguna akan perangkat pencarian dan informasi yang efektif dan efisien [1]. Efektif berarti pengguna mendapatkan dokumen yang relevan dengan query yang diinputkan. Efisien berarti juga hasil waktu pencarian yang lebih singkat.

Stemming merupakan proses yang memetakan bentuk varian kata menjadi kata dasarnya [1]. *Stemming* adalah inti dari suatu teknik pemrosesan natural language untuk mendapatkan informasi kembali (Information Retrieval) yang efektif dan efesien dan secara luas dapat diterima oleh pengguna. *Stemming* merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. Dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. *Stemming* juga dapat digunakan untuk mendukung proses kategorisasi/klasifikasi dan *clustering*. *Stemming* digunakan untuk mengganti bentuk/susunan dari suatu kata menjadi kata dasar dan dari kata tersebut yang telah sesuai dengan struktur morfologi Bahasa Indonesia yang baik dan benar.

Penggunaan *stemming* dalam bahasa Indonesia, memiliki dua jenis metode *stemming* yang sudah dikenal, yaitu *stemming* yang berbasis kamus (dictionary based) dan *stemming* yang berbasis non-kamus (purely rule based).

Algoritma *stemming* yang tidak menggunakan kamus memiliki tingkat kesalahan yang relatif tinggi, tapi di lain sisi algoritma tersebut memiliki kelebihan pada waktu proses yang lebih singkat dibandingkan algoritma *stemming* yang berbasis kamus. Kebanyakan algoritma *stemming* masih bergantung atau mengandalkan pada kamus untuk memeriksa apakah kata dasar dari sebuah kalimat ataupun kata yang telah dilakukan

proses *stemming* ditemukan atau tidak. Apabila kata dasar tersebut berhasil ditemukan pada kamus maka proses *stemming* dihentikan.

Imbuhan Bahasa Indonesia

Bahasa Indonesia selain memiliki imbuhan-imbuhan yang beraneka ragam, juga memiliki morfologi yang berbeda dengan bahasa lainnya. Sering kali sebuah kata dasar atau bentuk dasar perlu diberikan imbuhan agar dapat digunakan dalam pertuturan [2]. Imbuhan ini dapat mengubah makna, fungsi dan jenis sebuah kata dasar atau bentuk dasar menjadi kata lain yang fungsinya berbeda dengan kata dasar atau bentuk dasarnya.

Imbuhan mana yang harus digunakan tergantung pada keperluan penggunaannya di dalam pertuturan. Imbuhan yang ada dalam bahasa indonesia adalah :

1.Awalan (Prefix) ber-, per,-, me, di-, ter-, ke-, se-, dan pe-.Contoh : ber + jalan = berjalan, me + lompat = melompat, di + masak = dimasak, ter + jatuh = terjatuh, ke + toko = ketoko, se + kilo = sekilo, pe + maaf = pemaaf

2.Sisipan (Infix)-el-, -em, dan –er-.Contoh : telunjuk, jemari, gerigi

3.Akhiran (Suffix)-kan, -i, dan –nya.Contoh : bicara + kan = bicarakan, awal + i = awali, wujud + nya = wujudnya

4.Partikel (Particle)-lah, -kah, -pun.Contoh : pulang + lah = pulanglah, ada + kah = adakah, biar + pun = biarpun.

5.Kata ganti kepunyaan (Possive Pronoun)-ku, -mu, -nya.Contoh : sepeda + mu = sepedamu, sepeda+ ku = sepedaku, sepeda + nya = sepedanya.

6.Imbuhan gabung (Confix)ber-kan, ber-an, per-kan, per-i, me-kan, me-i, memper-, memper-kan, memper-i, di-kan, di-i, diper-, diper-kan, diper-i.Contoh : ber + alas + kan = beralaskan, ber + dua + an = berduaan, per + bincang + kan = perbincangkan, memper + banding + kan = memperbandingkan.

Algoritma*Stemming* Porter

Porter Stemmer for Bahasa Indonesia berdasarkan English Porter Stemmer yang dikembangkan oleh W.B. Frakes pada tahun 1992. Karena bahasa Inggris datang dari kelas yang berbeda, beberapa modifikasi telah dilakukan untuk membuat Algoritma Porter dapat digunakan sesuai dengan bahasa Indonesia[1].

Tahapan dan Langkahnya:

1. Hapus Particle.
2. Hapus Possesive Pronoun.
3. Hapus awalan pertama.

Jika tidak ada lanjutkan ke langkah 4a, jika ada carimaka lanjutkan ke langkah 4b.

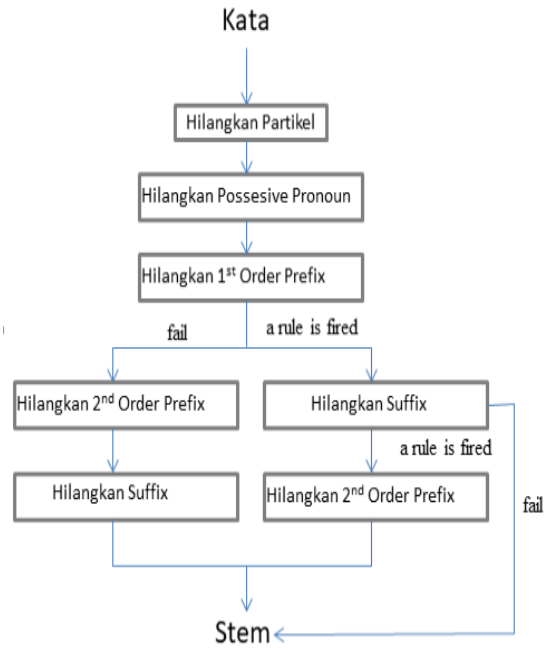
4.a. Hapus awalan kedua, lanjutkan ke langkah 5a.

4.b. Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai root word.

Jika ditemukan maka lanjutkan ke langkah 5b.

5.a. Hapus akhiran. Kemudian kata akhir diasumsikan sebagai root word.

5.b.Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai root word.



Gambar 1.*Alur proses stemming Porter Bahasa Indonesia*[1]

Algoritma *Stemming*Yusoff Sembok

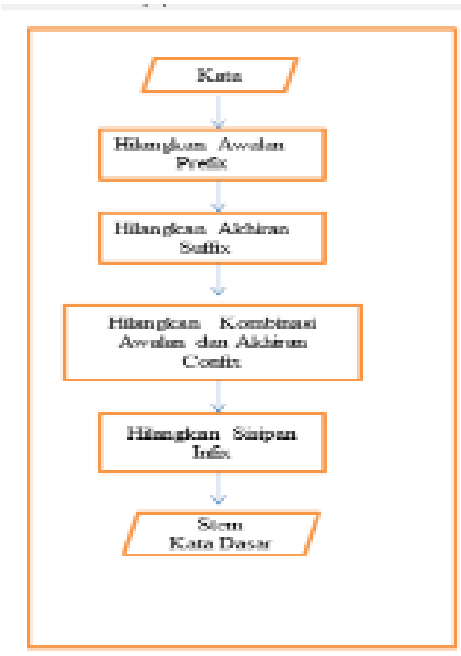
Algoritma Ahmad Yusoff Sembok dikembangkan sebagai pendekatan baru dalam *stemming*. Stemmer ini tidak sebaik stemmer lain namun ada beberapa kata yang tidak dapat diproses pada stemmer lain seperti Nazief &Adriani namun dapat diproses oleh Ahmad Yusoff Sembok.

Algoritma Ahmad Yussof dan Sembok tidak memiliki standar dan aturan yang khusus terhadap urutan pemotongan imbuhan pada tiap-tiap katanya. Algoritma Ahmad Yusoff dan Sembok ini memiliki beberapa variasi urutan (orders) pemotongan imbuhan,dimana imbuhan tersebut terdiri dari 4 jenis, yaitu:

1.Prefixes(melakukan pemotongan Awalan)
2.Infixes (melakukan pemotongan Sisipan)
3.Confixes(Gabungan /Kombinasi awalan dan akhiran)
4.Suffixes(melakukan pemotongan akhiran)

Dari beberapa variasi urutan pemotongan imbuhan pada algoritma Ahmad Yusoff dan Sembok, maka telah disimpulkan bahwa urutan yang paling efektif menurut[2] adalah:

1.Prefixes(melakukan pemotongan Awalan)
2.Suffixes(melakukan pemotongan akhiran)
3.Confixes(Gabungan /Kombinasi awalan dan akhiran)
4.Infixes(melakukan pemotongan Sisipan)



Gambar 2.*Alur proses stemming Ahmad Yusoff Sembok* [2]

Stemming Berbahasa Indonesia, Ada beberapa teknik pendekatan seperti pendekatan *stemming* yang dilakukan oleh stemmer Porter, Tala, Vega, Arifin dan Setiono, Nazief dan Adriani[4]. Sebenarnya hampir tidak ada persetujuan umum yang bersifat baku mengenai keefektifan dari teknik-teknik pendekatan tersebut. Masalah lainnya dari *stemming* adalah masih bergantung dari beberapa teknik *stemming* tersebut pada kamus yang lebih luas (comprehensive dictionary). Dalam penelitian ini akan dilakukan pengukuran efektifitas dari 2 (dua) algoritma yang digunakan pada proses *stemming* baik yang menggunakan kamus yaitu algotima *stemming* porter maupun menggunakan aturan imbuhan yaitu algoritma *stemming* yusoff sembok. Evaluasi pengukuran dilakukan pada tingkat keakuratan, *overstemming*dan *understemming.*

Selain hanya bisa menentukan hasil keakuratan, *overstemming* dan *understemming*algoritma *stemming* juga bisa membantu mengkalsifikasikan model dokumen kalimat berupa teks yaitu dengan mengukur akurasi[5]

METODE PENELITIAN

Metode penelitian yang dilakukan dalam penelitian ini adalah dengan melakukan analisa objek penelitian, desain penelitian dan teknik pengumpulan data yang dilakukan dari berbagai sumber, selanjutnya metode yang diusulkan. Kemudian melakukan preprosesing data dan penerapan algoritma. Langkah selanjutnya yang akan dilakukan adalah melakukan eksperimen/pengujian terhadap data set pada masing – masing algoritma yang digunakan.

Evaluasi pengukuran

Pada penelitian ini akan dilakukan sebuah evaluasi terhadap masing-masing hasil *stemming* pada setiap dokumen yang ada. Untuk pengukurannya dilakukan secara detail mulai dari dokumen pertama sampai dokumen yang terakhir. Masing-masing algoritma *stemming* akan diujikan menggunakan 100 dokumen. Adapun pengukuran yang dilakukan adalah sebagai berikut :

1.Akurasi

Akurasi hasil pengujian didapatkan dari perbandingan hasil *stemming* dengan relevance judgments yang telah dibuat, yang hasilnya dibagi dengan jumlah kata dalam dokumen.

Akurasi =
(Kata dasar hasil *stemming* = Relevance judgments) x 100%
Jumlah kata dalam dokumen
[1]

2.*Overstemming*adalah kata-kata yang terlalu banyak dipotong setelah dilakukan proses *stemming* dibandingkan dengan *relevance judgments*[1].

3.*Understemming* adalah kata yang terlalu sedikit dipotong setelah dilakukan proses *stemming* dibandingkan dengan *relevance judgments*[1].

Desain Tahapan Penelitian

Penelitian ini merupakan penelitian eksperimen dengan tahapan metode penelitian sebagai berikut:

1. Objek penelitian

Obyek penelitian dalam penelitian ini adalah kata-kata bahasa Indonesia baik kata dasar maupun kata berimbuhan yang berupa kalimat dan kata dalam bentuk dokumen.

2.Pengumpulan Data

Tahapan ini akan dijelaskan bagaimana dan darimana data digunakan untuk penelitian ini didapatkan, data dalam penelitian ini diambil dari berbagai sumber dan dijadikan sebagai dokumen uji dengan ekstensi.txt.

3. Metode yang Diusulkan

Metode ini akan menjelaskan tentang pendekatan yang diusulkan untuk melakukan proses *stemming*, baik menggunakan kamus dan aturan imbuhan.

4.Preprocessing (*Case folding* dan *Tokenizing*)

Tahapan ini akan dijelaskan tentang tahap awal untuk melakukan ekstraksi sebuah dokumen sebelum dilakukan proses *stemming*, yaitu dengan menghilangkan tanda spasi, karakter dan angka dalam dokumen. Selanjutnya akan dilakukan pemisahan kalimat menjadi per- kata dalam dokumen.

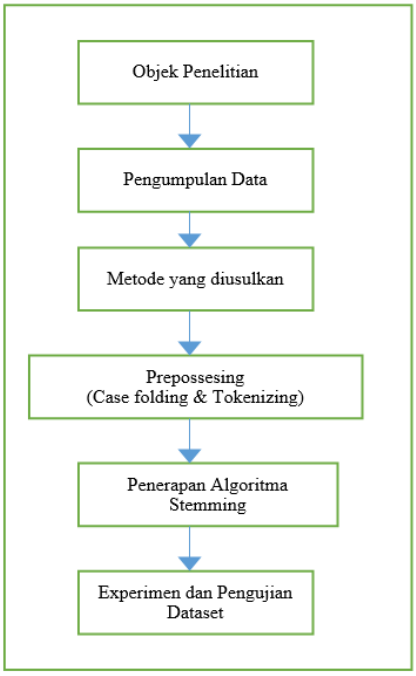
5.Penerapan Algoritma *Stemming*

Tahap ini akan dijelaskan tentang langkah - langkah untuk menerapkan algoritma *stemming*, sehingga dapat dibuktikan bahwa pemilihan metode ini bisa digunakan untuk menghasilkan kata dasar dari sebuah kata dalam dokumen terutama kata berimbuhan.

6.Eksperimen dan Pengujian data set

Bagian ini dijelaskan tentang langkah-langkah eksperimen yang meliputi cara desain eksperimen sampai memperoleh hasil kata dasar dari sebuah dokumen yang diujikan / di*stemming*. Hasil eksperimen akan menunjukkan bahwa metode yang digunakan adalah tepat.

Desain tahapan penelitian ini diperlukan untuk tahapan awal sebelum proses ujicoba dilakukan sehingga nanti pada tahapan eksperimen bisa dibandingkan antara kedua pendekatan algoritma*stemming*perter dengan algoritma*stemming*yusoff sembok.
Setiap eksperimen algoritma yang digunakan dan dibandingkan akan memberikan tentunya hasil *stemming*yang berbeda dan *undesstemming* yang berbeda pula, berikut adalah gambar metode tahapan penelitian.

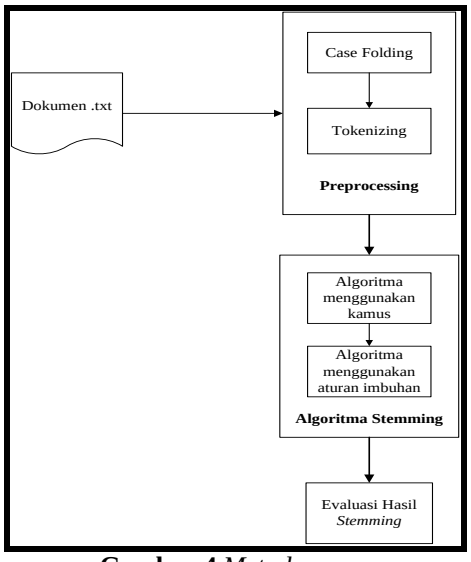


Gambar 3.*Tahapan Metode Penelitian(penulis)*

Metode yang diusulkan

Metode yang diusulkan untuk penelitian ini menggunakan

pendekatan algoritma *stemming* berbasis kamus dan menggunakan aturan imbuhan untuk dokumen teks Bahasa Indonesia. algoritma yang digunakan dalam penelitian ini adalah algoritma Porter dan algoritma Ahmad Yusoff Sembok.



Gambar 4.*Metode yang diusulkan*(penulis)

Tahap *Preprocessing*

Tahap ini menjelaskan tentang langkah preprocessing pada algoritma *stemming*. Langkah preprocessing ini digunakan untuk menghilangkan bagian-bagian yang tidak diperlukan yang terdapat pada sebuah dokumen dimana hal ini akan menjadi noise pada proses selanjutnya, selain itu langkah preprocessing ini sendiri berfungsi sebagai parameter input algoritma *stemming*. *Preprocessing* yang dilakukan ini hanya pada tahapan *case folding* dan *tokenizing* saja.

Case Folding dan Tokenizing

Case folding adalah merupakan suatu proses mengubah dokumen / kalimat menjadi huruf kecil, namun yang dirubah hanya huruf saja bukan angka maupun simbol-simbol lainnya. Selain itu *Case folding* juga merupakan proses penghapusan spasi, angka dan tanda baca dalam dokumen / kalimat.

Foto Cagub Diusulkan Dipasang di Kantor Pemerintah Untuk menekan angka golongan putih (golput) pada Pemilihan Gubernur (Pilgub) DKI putaran kedua,

Gambar 5. *Contoh kalimat dalam sebuah dokumen(penulis)*

foto cagub diusulkan dipasang di kantor pemerintah untuk menekan angka golongan putih golput pada pemilihan gubernur pilgub dki putaran kedua

Gambar 6. *Setelah dilakukan proses case folding(penulis)*

Tokenizing adalah suatu proses pemotongan /pemenggalan kalimat berdasarkan pada kata-kata yang menyusunnya.

foto cagub diusulkan dipasang di kantor pemerintah untuk menekan angka golongan putih golput pada pemilihan gubernur pilgub dki putaran kedua

Gambar 7.*Contoh kalimat dalam sebuah dokumen(penulis)*

Setelah dilakukan proses *tokenizing* hasilnya menjadi :

foto	golong
cagub	putih
usul	golput
pasang	pada
di	pilih
kantor	gubernur
pemerintah	pilgub
untuk	dki
tekan	putar

Gambar 8.*Contoh proses tokenizing(penulis)*

Tahap ini sangat penting dilakukan karena akan mempengaruhi hasil inputan kata dalam dokumen yang selanjutnya akan dilakukan proses *stemming*.

Algoritma *Stemming*

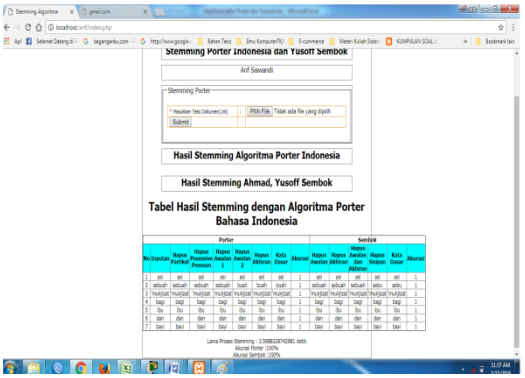
Proses *stemming* akan menghasilkan kata dasar dari setiap kata hasil *tokenizing* dalam setiap dokumen. Dalam penelitian ini akan digunakan algoritma Porter Indonesia dan algoritma Ahmad Yussof Sembok untuk mendapatkan kata dasar dalam bahasa Indonesia. Dalam proses ini dilakukan pemotongan terhadap imbuhan sehingga dihasilkan kata dasar dari kata tersebut. Masing - masing algoritma akan diujikan dengan dataset yang sama dan hasilnya akan dievaluasi.

Eksperimen dan pengujian dataset

Sebelum diterapkan dan diujikan pada dataset berupa dokumen sampel yang sudah ditentukan sebelumnya, maka model perlu diujikan terlebih dahulu dengan menggunakan dataset dokumen yang sudah dalam bentuk .txt. Dokumen yang sudah disediakan sebagai bahan uji berjumlah 100 dokumen dengan jumlah kata yang bervariasi disetiap dokumen. Kamus yang digunakan pada pengujian adalah kamus yang berisikan kata dasar saja. Hasil *stemming* kemudian akan dibandingkan dengan relevance judgment yang dibuat berdasarkan pengetahuan manusia.

Untuk memudahkan dalam melakukan pengujian maka dilakukan testing model terhadap aplikasi *stemming* yang sudah dibuat sebelumnya. Hasil pengujian dari masing-masing dokumen yang sudah di*stemming* akan menunjukkan jumlah kata dalam dokumen,dan juga kata dasar hasil *stemming*. Pengujian ini dilakukan pada aplikasi yang sudah dibuat untuk memperoleh hasi *stemming*. Proses selanjutnya adalah mengevaluasi terhadap hasil *stemming*. Evaluasi dilakukan untuk mengetahui nilai akurasi dari masing-masing algoritma dan juga kata-kata yang mengalami *overstemming* dan *understemming*.

Gambar 9.*Proses input dokumen(penulis)*



Gambar 10.*Hasil Input(penulis)*

Kelebihan dan Kekurangan algoritma *stemming*

Dari sekian banyak algoritma yang digunakan untuk melakukan proses *stemming*, khususnya *stemming* bahasa Indonesia. Masing-masing algoritma memiliki kelebihan dan kekurangan , berikut penjelasannya.

Algoritma Porter indonesia

Kelebihan :

1.Memperhatikan kemungkinan adanya partikel-partikel yang mungkin mengikuti suatu kata berimbuhan.

2.Algoritma ini mengacu pada aturan morfologi bahasa Indonesia yang mengelompokkan imbuhan, yaitu imbuhan yang diperbolehkan atau imbuhan yang tidak diperbolehkan.

3.Algoritma ini menggunakan kamus kata dasar untuk meningkatkan keakuratan hasil *stemming*.

Kekurangan :

1.Penyamarataan makna variasi kata.

2.Kesalahan terjadi bila kata tidak ditemukan di database dan kemudian dianggap kata dasar.

3.Membutuhkan waktu untuk proses pencarian kata di dalam kamus.

Algoritma Ahmad Yusoof Sembok

Kelebihan :

1.Jika kata dasar dari sebuah kata turunan tidak dapat ditemukan setelah menghilangkan prefix dan suffix nya maka algoritma ini akan mencoba mengembalikan semua imbuhan yang telah dihilangkan tadi untuk dikombinasikan dengan kata hasil *stemming* dari kata turunan yang kata dasarnya tidak ditemukan pada kamus.

2.Algoritma ini menggunakan kamus kata dasar untuk membantu meningkatkan keakuratan hasil *stemming*.

Kekurangan :

1.Waktu proses yang lebih lama karena ada pengecekan setiap langkah terhadap kamus dalam database.

2.Penyamarataan makna variasi kata.

HASIL DAN PEMBAHASAN

Di dalam bagian ini pembahasan akan difokuskan mengenai hasil percobaan yang dilakukan mulai dari koleksi dokumen yang digunakan, kamus kata dasar, serta relevance judgment untuk membandingkan kata dasar hasil *stemming* dengan kata dasar menurut pengetahuan manusia. Berdasarkan hasil penelitian yang telah dilakukan di dapat hasil yang berbeda antara algoritma yang satu dengan yang lainnya. Pengujian ini dilakukan masing-masing 100 kali untuk setiap algoritma. Untuk mengetahui tingkat performansi dari masing-masing algoritma maka akan dilakukan pengukuran terhadap nilai akurasi kata, kata-kata yang *overstemming*dan *understemming.*

Untuk koleksi dokumen yang digunakan dalam pengujian adalah dokumen sampel sebanyak 100 dokumen, yang sudah dibuat menggunakan format berekstensi .txt. Kata-kata dalam dokumen ini diperoleh dari isi artikel-artikel dan berita, baik berita dan artikel tentang teknik, berita dan artikel tentang kesehatan, artikel keilmuan maupun berita-berita dari media elektronik. Total kata-kata secara keseluruhan pada 100 dokumen adalah 25.819 kata. Isi kata dalam setiap dokumen bervariasi mulai dari puluhan kata sampai ribuan kata, dimana kata-kata di dalam dokumen tersebut belum dilakukan pengolahan teks atau preprocessing.

Dari hasil proses *stemming* yang pernah dilakukan sebelumnya oleh Nazief dan Adriani, serta Arifin dan Setiono, dapat di buat kesimpulan bahwa diperlukan sebuah kamus kata dasar untuk mendapatkan hasil *stemming* yang baik. Kamus kata dasar tersebut diperlukan dan dibutuhkan untuk memeriksa hasil dari proses *stemming* apakah kata dasar yang melalui proses *stemming* ini benar dan sesuai serta ditemukan pada kamus kata dasar yang berada didalam database saat proses *stemming* dilakukan. Semakin lengkap kamus kata dasar yang digunakan maka nilai akurasi hasil *stemming* akan semakin besar. Pada penelitian ini, kamus kata dasar diambil dari daftar kata dasar pada Kamus Besar Bahasa Indonesia (KBBI) luring CHM V1.5 yang diunduh dari ebsoft.web.id. Total kata dasar dalam kamus adalah sebanyak 31.295 kata dasar.

Pada penelitian ini, sebelum dilakukan evaluasi tentang performa hasil *stemming* baik akurasi kata maupun kata-kata yang mengalami *overstemming* dan *understemming* dalam setiap dokumen maka dibuat relevance judgments per kata untuk melihat seberapa relevannya hasil *stemming* terhadap relevance judgments yang dibuat berdasarkan pengetahuan manusia. Relevance judgments yang dibuat adalah hanya berupa kata dasar saja.

Pengetahuan kita sebagai manusia mengenai kata dasar yang baik dan benar sangat dibutuhkan, karena semakin tinggi tingkat pengetahuan manusia mengenai kata dasar dan juga sebuah kata berimbuhan maka hasilnya akan semakin baik. Hal ini dilakukan untuk membandingkan antara kata dasar hasil *stemming* yang dilakukan dengan menggunakan komputer dan kata dasar yang dihasilkan berdasarkan pengetahuan manusia.

Tabel 1.*Relevance judgments pada dokumen(penulis)*

No	Inputan Kata	Hasil *Stemming* (Kata Dasar)	*Relevance Judgments* (Kata Dasar)
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

Hasil pengujian 100 dokumen terhadap akurasi hasil *stemming*.

Hasil pengujian yang dilakukan terhadap nilai dari proses akurasi *stemming* menunjukkan bahwa rata-rata nilai akurasi yang paling besar terdapat pada algoritma *stemming* Porter.

Tabel 2. *Rata-rata akurasi hasil stemming(penulis)*

Jumlah dokumen (100)	Algoritma Porter	Algoritma Ahmad, Yusoof Sembok
Rata-rata akurasi (%)	94,470	86,135

Hasil pengujian 100 dokumen terhadap *overstemming*

Hasil pengujian yang dilakukan terhadap nilai dari proses *overstemming* menunjukkan bahwa rata-rata *overstemming* yang paling kecil terdapat pada algoritma *stemming* Porter.

Tabel 3.*Rata-rata presentase overstemming(penulis)*

Jumlah dokumen (100)	Algoritma Porter	Algoritma Ahmad, Yusoof Sembok
Rata-rata *overstemming* (%)	4,541	5,023

Hasil pengujian 100 dokumen terhadap *understemming*

Hasil pengujian yang dilakukan terhadap nilai dari proses *understemming*menunjukkan bahwa rata-rata *understemming* yang paling kecil terdapat pada algoritma *stemming* Porter.

Tabel4.*Rata rata presentas eunderstemming(penulis)*

Jumlah dokumen (100)	Algoritma Porter	Algoritma Ahmad, Yusoof Sembook
Rata-rata *understemming* (%)	0,989	8,867

Analisa kesalahan hasil proses *stemming*

Evaluasi terhadap kesalahan hasil *stemming* pada setiap algoritma hasilnya berbeda-beda, kesalahan terjadi karena dipengaruhi masing-masing algoritma yang memiliki karakteristik yang berbeda. Ketika sebuah kata dilakukan proses *stemming*. Kesalahan tidak hanya terjadi pada kata dasar dan kata berimbuhan saja, tetapi kesalahan juga disebabkan karena salah pengetikan kata yang ditemukan dalam dokumen yang digunakan. Selain itu kesalahan juga ditemukan dan terjadi pada nama orang, nama tempat dan istilah termasuk bahasa asing, baik pada algoritma yang menggunakan kamus maupun pada algoritma yang menggunakan aturan imbuhan. Berikut kesalahan hasil *stemming* karena pengetikan kata yang salah didalam dokumen yang digunakan.

Tabel 5.*Kesalahan hasil stemming karena kesalahan pada pengetikan kata(penulis)*

Contoh	Seharusnya	Hasil *stemming*	Kata dasar
adaalah	adalah	adaalah	adalah
bentuan	bantuan	bentu	bantu
bartahun	bertahun	bartahun	tahun
denganya	dengannya	denga	dengan
dikembamngkan	dikembangkan	dikembamngkan	kembang
diataati	ditaati	diataati	taat
inginya	inginnya	ingi	ingin
kuargaku	keluargaku	kuarga	keluarga
kemorosotan	kemerosotan	kemorosotan	merosot
kemajuanm	kemajuan	kemajuanm	maju

keberlangssungan	keberlangsungan	langssung	langsung
kemudia	kemudian	kemudia	kemudian
lagkah	langkah	lag	langkah
menyadihkan	menyedihkan	menyadihkan	sedih
mengispirasi	menginspirasi	mengispirasi	inspiasi
menyanyakan	menanyakan	sanya	tanya
menyutujui	menyetujui	menyutujui	setuju
menganggab	menganggap	menganggab	anggap
menyebrangi	menyeberangi	menyebrangi	seberang
menyipan	menyimpan	sip	simpan
menyelediki	menyelidiki	menyelediki	selidik
peghiajaua	penghijauan	penghijaua	hijau
pesanya	pesannya	pes	pesan
tetang	tentang	tang	tentang
tebuat	terbuat	tebuat	buat

Berikut kesalahan hasil *stemming* terhadap nama orang, nama tempat dan istilah

Tabel 6. *Kesalahan hasil stemming terhadap nama orang, tempat dan istilah.(penulis)*

Contoh	Hasil *stemming*	Seharusnya
*abdullah	abdul	tetap
asian	asi	tetap
bakri	bakr	tetap
bisri	bisr	tetap
budhiman	budhim	tetap
chasbullah	chasbul	tetap
dahlan	dahl	tetap
diego	ego	tetap
edi	ed	tetap
fauzi	fauz	tetap
fahri	fahr	tetap
jokowi	jokow	tetap
panjaitan	panjait	tetap
robi	rob	tetap
syamsuri	syamsur	tetap
setiono	tiono	tetap
vani	van	tetap
wahyudi	wahyud	tetap
yadi	yad	tetap
zuhri	zuhr	tetap
*betawi	betaw	tetap
bali	bal	tetap
bekasi	bekas	tetap
dki	dk	tetap
indian	indi	tetap
kalimantan	kalimant	tetap
kudus	dus	tetap
sulawesi	sulawes	tetap
*deputi	deput	tetap
diet	et	tetap
fireman	firem	tetap
indrawi	indraw	tetap
bayi	bay	tetap
kyai	kya	tetap
lan	l	tetap
pemadam	padam	tetap
pengawa	tawas	tetap

Berikut ditampilkan kesalahan hasil *stemming* terhadap bahasa asing dapat dilihat

Tabel 7. *Kesalahan hasil* stemming *terhadap bahasa asing(penulis)*

Contoh	Hasil *stemming*	Seharusnya
fireman	firem	tetap
medicine	cine	tetap
members	s	tetap
megaphone	gaphone	tetap

Analisa kesalahan hasil* stemming *pada algoritma Porter.

Untuk mengetahui ketepatan hasil *stemming* perlu dilakukan analisa secara manual. Mengingat jumlah kata yang cukup besar (25.819 kata), pengamatan mencakup sebagian saja. Hasil analisa diambil dari kata-kata yang gagal di *stemming* yang sudah disortir sebelumnya. Kesalahan hasil *stemming* pada algoritma Porter adalah apabila kata tidak ditemukan di kamus database dan kemudian dianggap kata dasar. Berikut kesalahan hasil *stemming* pada algoritma Porter terhadap kata berimbuhan.

Tabel 8.*Kesalahan hasil stemming pada algoritma Porter.(penulis)*

Contoh	Hasil *stemming*	Seharusnya
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
berupa	upa	Rupa
bukanlah	bu	Bukan
dariku	dar	Dari
denganya	denga	Dengan
diadakan	adakan	Ada

dianjurkan	dianjurkan	Anjur
diataati	diataati	Taat
didiknya	dik	Didik
dikemukakan	dikemukakan	Muka
diketahui	diketahui	Tahu
dirediksikan	dirediksikan	Rediksi
dirinya	r	Diri
disahkan	sahkan	Sah
ditandatangani	ditandatangani	Tandatangan
ditelah	dite	Ditelah
gerakannya	gera	Gerak
kalinya	kal	Kali
keberagaman	agam	Ragam
kebijakan	bija	Bijak
kekerasan	kerasan	Keras
kelap	lap	Kelap
kelasnya	las	Kelas
kemorosotan	kemorosotan	Morosot
kepalanya	pala	Kepala

Analisa kesalahan hasil *stemming* pada algoritma Ahmad Yussof Sembok

Kesalahan yang terjadi pada algoritma Ahmad Yussof Sembok disebabkan karena aturan pemotongan awalan, akhiran, sisipan dan kombinasi awalan dan akhiran. Contoh kata *"pemuda"* dalam algoritma tersebut dianggap memiliki awalan pe- padahal sebenarnya adalah kata dasar, sehingga setelah dilakukan pemotongan dan hasil *stemming* menjadi *"uda"*. Kesalahan juga terjadi pada kata-kata berimbuhan yang tidak mengalami perubahan kata sebelum dilakukan *stemming* dan sesudah dilakukan *stemming* khususnya pada kata majemuk, contohnya kata "berterimakasih" setelah dilakukan *stemming* hasilnya tetap "berterimakasih". Untuk lebih jelasnya kesalahan hasil *stemming* dalam algoritma Ahmad Yussof Sembok.

Tabel 9. *Kesalahan hasil stemming pada algoritma Ahmad Yusoff Sembok(penulis)*

Contoh	**Hasil *stemming***	**Seharusnya**
bekerjasama	bekerjasama	Kerjasama
berdimensi	mens	Dimensi
berdiri	r	Diri
bereaksi	bereaksi	Reaksi
berekpresi	berekpresi	Ekspresi
berikan	ikan	Beri
berisiko	berisiko	Risiko
berkelindan	berkelindan	Kelindan
berkembangnya	berkembangnya	Kembang
berkurangnya	rang	Kurang
berpengaruh	taruh	Pengaruh
berpesan	s	Pesan
bersekolah	kolah	Sekolah
berselang	lang	Selang
bersepeda	peda	Sepeda
bertanggungjawab	bertanggungjawab	Tanggungjawab
berterimakasih	berterimakasih	Terimakasih
bertopologi	bertopologi	Topologi
berupa	upa	Rupa
beserta	beserta	Serta
dia	a	Dia
diberikan	ikan	Beri
didik	dik	Didik
didiknya	dik	Didik

Analisa algoritma *stemming* Porter dengan algoritma Ahmad Yussof Sembok.

Dalam pengujian yang telah dilakukan diperoleh hasil akurasi yang paling baik yaitu terdapat pada algoritma Porter dibandingkan dengan algoritma Ahmad Yussof sembok untuk algoritma yang menggunakan kamus. Yang menjadi perbedaan diantara kedua algoritma tersebut adalah aturan yang dipergunakan masing-masing algoritma. Algoritma Porter terdapat imbuhan yang diperbolehkan dan imbuhan yang tidak diperbolehkan. Algoritma Porter mengusulkan penambahan aturan-aturan seperti penambahan aturan untuk reduplikasi dan penambahan aturan untuk awalan dan akhiran, untuk meningkatkan presisi dari setiap kata yang di *stemming*. Sedangkan algoritma Ahmad Yussof sembok membuat aturan pemotongan awalan, akhiran, sisipan dan kombinasi awalan dan akhiran. Perbedaan kedua algoritma dalam mengeksekusi sebuah kata.

Tabel 10.*Tabel perbedaan hasil stemming algoritma Porter dengan algoritma Ahmad Yusoff Sembok(penulis)*

Algoritma	**Inputan Kata**	**Imbuhan**	**Hasil *stemming***	**Seharusnya**
	memberikan	mem-, -kan	beri	sudah benar

Porter	berkurang	ber-	kurang	sudah benar
	nilainya	-nya	nila	Nilai
	kepada	-	kepada	sudah benar
	berupa	-	upa	Rupa
	kepentingan	ke-, -an	penting	sudah benar
	kerusakan	ke-, -an	rusa	Rusak
	tersebut	ter-	sebut	sudah benar
	sedikit	-	sedikit	sudah benar
	mengurangi	meng-, -i	urang	Kurang
	menipis	men-	nipis	Tipis
	sejumlah	se-, -lah	sejum	Sejumlah
	sesuai	-	sesuai	sudah benar
Ahmad yussof & Sembok	memberikan	mem-, -kan	pi	Beri
	berkurang	ber-	rang	Kurang
	nilainya	-nya	nilai	sudah benar
	kepada	-	pada	Kepada
	berupa	-	upa	Rupa
	kepentingan	ke-, -an	ting	Penting
	kerusakan	ke-, -an	rusa	Rusak
	tersebut	ter	terbut	Sebut
	sedikit	-	dikit	Sedikit
	mengurangi	meng-, -i	kurang	Kurang
	menipis	men-	tipis	Tipis
	sejumlah	se-, -lah	sejum	Sejumlah
	sesuai	-	suai	Sesuai

PENUTUP

Kesimpulan

1.Dengan menggunakan algoritma *stemming* porter berbahasa Indonesia terbukti memiliki tingkat akurasi yang baik.

2.Kesalahan *stemming* porter pada tahapan proses awal sebelum proses terjadi dimana kata tidak ditemukan di kamus database dan kemudian dianggap kata dasar.

3.Kesalahan analisis terjadi karena adanya kata yang sama muncul pada Beberapa dokumen yang berbeda.

Saran

1. Melakukan analisa lainya dengan perbandingan beberapa Algoritma *stemming*
2. Untuk mengurangi tingkat kesalahan dalam seleksi kata untuk kata dasar sebaiknya disesuaikan terlebih dahulu dengan menggunakan KKBI dan tentukan fitur atributnya, dilakukan penelitian dengan algoritma *stemming* yang berbeda untuk hasil *case* yang lebih baik ke depannya.

DAFTAR PUSTAKA

[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] Chaer, A 2011. Tata Bahasa Praktis Bahasa Indonesia, Rineka Cipta Jakarta. Indonesia.

[3] Asian, Jelita. "Effective techniques for Indonesian

[4] Simarangkir, ManaseSahat H 2017.Studi perbandingan algoritma - algoritma stemming untuk dokumen teks bahasa indonesia. Jurnal Inkofar Politeknik meta industri.

[5] 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