Stemming teks *sor-singgih* Bahasa Bali

Gusti Ngurah Mega Nata[1)], Putu Pande Yudiastra[2)]
STMIK STIKOM Bali[1,2]
Jl. Raya Puputan No.86 Renon, telp: (0361) 244445/fax. (0361) 264773
e-mail: mega@stikom-bali.ac.id[1)], yudiastra87@gmail.com[2)]

Abstrak

Bahasa Bali memiliki tingkatan penggunaanya yaitu Bali Alus, Bali Madya dan Bali Kasar yang lebih dikenal dengan sor-singgih Bahasa Bali. Text mining menggunakan dokumen Bahasa Bali merupakan suatu tantangan karena sor-singgih Bahasa Bali tersebut. Sor-singih Bahasa Bali menyebabkan masalah pada hasil stemming, karena setiap level Bahasa memiliki kata dasar sendiri – sendiri tetapi memiliki sematik yang sama. Sehingga, dimensi kata dasar akan menjadi sangat banyak pada saat proses text mining. Pada penelitian ini, akan dikembangkan algoritma stemming khusus untuk Bahasa Bali dalam menanggani sor-singgih pada dukumen bahasa Bali. Pada proses stemming algoritma yang akan digunakan yaitu algoritma Porter Stemmer for Bahasa Indonesia yang disesuaikan kembali dengan morfologi dan Afiks dari bahasa Bali dan juga akan disediakan padanan kata dasar dari Bahasa Bali sor-singgih dengan bahasa Indonesia. Dari hasil pengujian 85% kata distemming dengan benar. Hasil dari stemming yang berupa kata dasar kemudian ditranslate / dicarikan padanannya dengan bahasa Indonesia.

Kata kunci: *stemming, kata dasar, Bahasa Bali, sor-singgih, dokumen*

1. Pendahuluan

Bahasa Bali merupakan bahasa Austronesia dari cabang sundik dan lebih spesifik dari anak cabang Bali-Sasak. Bahasa Bali memiliki tingkatan penggunaanya yaitu Bali Alus, Bali Madya dan Bali Kasar yang lebih dikenal dengan *sor-singgih* Bahasa Bali [1]. Bahasa Bali Alus digunakan untuk berbicara formal atau untuk orang berkasta lebih tinggi, Bahasa Bali Madya digunakan bagi golongan menengah yang tidak terlalu formal, sedangkan untuk Bahasa Bali kasar digunakan untuk kalangan sudra atau golongan berkasta rendah. Walaupun memiliki perbedaan ucapan pada setiap Bahasa Bali namun, Pada dasarnya memiliki simantik makna yang sama, seperti Bahasa alus dari kata "saya" yaitu "*titiang",* dengan bahasa kasar: *"icang",* atau bahasa alus "berjalan" yaitu "*mamargi"* kata dasarnya "*margi*" dengan bahasa Madya berjalan : "*mejalan"* kata dasarnya "*jalan*".

Text Mining menggunakan dokumen bahasa Bali sudah mulai dilakukan pada paper [2] namun, tidak membahas *sor-singgih* Bahasa Bali. Padahal sor-singgih Bahasa Bali merupakan suatu tantangan pada saat melakukan *text mining* dimana, Bahasa tersebut memiliki tingkatan Bahasa namun memiliki semantic yang sama. Satu dokumen bahasa Bali bisa mengandung satu atau lebih tingkatan bahasa / *sor-singgih*. *Sor-singih* bahasa Bali menyebabkan masalah pada hasil *stemming*, karena setiap level bahasa Bali memiliki kata dasar sendiri – sendiri tetapi memiliki semantik yang sama. Sehingga, dimensi kata dasar akan menjadi sangat banyak pada saat proses *text mining*. Selain itu, proses pengelompokan (*classtering)* dokumen sangaat dipengaruhi oleh tingkatan Bahasa yang digunakan pada setiap dokumen tersebut. Jadi, Dalam dokumen Bahasa Bali proses *stemming* untuk setiap level bahasa Bali harus merujuk ke satu kata dasar untuk menggurangi dimensi kata.

Pada proses *stemming* algoritma yang akan digunakan yaitu algoritma *Porter Stemmer* for Bahasa Indonesia yang dikembangkan oleh Fadillah Z Tala pada tahun 2003 [4] yang disesuaikan kembali dengan morfologi dan Afiks dari bahasa Bali. Algoritma porter Bahasa Indonesia memiliki tingkat kecepatan yang paling baik dari algoritma Bahasa Indonesia lainnya seperti algoritma CS [5,6]. Hasil *stemming* yang berupa kata dasar kemudian akan di*translate* ke kata dasar Bahasa Indonesia untuk mengurangi dimensi kata dasar Bahasa Bali. *Translate* kata dasar berupa *list of word* Bahasa Bali dan Bahasa Indonesia yang hanya untuk menemukan padanan kata dari satu tingkatan Bahasa Bali ke Bahasa Indonesia.

2. Metode Penelitian

2.1 Studi Literatur

1) Bahasa Bali

Bahasa Bali adalah sebuah bahasa Austronesia dari cabang Sundik dan lebih spesifik dari anak cabang Bali-Sasak. Bahasa ini terutama dipertuturkan di pulau Bali, pulau Lombok bagian barat, dan sedikit di ujung timur pulau Jawa. Di Bali sendiri Bahasa Bali memiliki tingkatan penggunaannya, misalnya ada yang disebut Bali Alus, Bali Madya dan Bali Kasar [1]. Yang halus dipergunakan untuk bertutur formal misalnya dalam pertemuan di tingkat desa adat, meminang wanita, atau antara orang berkasta rendah dengan berkasta lebih tinggi. Yang madya dipergunakan di tingkat masyarakat menengah misalnya pejabat dengan bawahannya, sedangkan yang kasar dipergunakan bertutur oleh orang kelas rendah misalnya kaum sudra atau antara bangsawan dengan abdi dalemnya, Di Lombok bahasa Bali terutama dipertuturkan di sekitar kota Mataram, sedangkan di pulau Jawa bahasa Bali terutama dipertuturkan di beberapa desa di kabupaten Banyuwangi. Selain itu bahasa Osing, sebuah dialek Jawa khas Banyuwangi, juga menyerap banyak kata-kata Bali. Misalkan sebagai contoh kata *osing* yang berarti "tidak" diambil dari bahasa Bali *tusing*. Bahasa Bali dipertuturkan oleh kurang lebih 4 juta jiwa.

2) *Stemming*

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya [2]. Proses *stemming* untuk setiap Bahasa berbeda dengan Bahasa yang lain misal, proses *stemming* Bahasa Inggris dengan Bahasa Indonesia tentunya berbeda karena perbedaan pembentukan dan perubahan kata menjadi bentuk kata lain [6]. Dalam dokumen Bahasa Indonesia proses *stemming* sangat diperlukan sebelum proses *text mining* karena Bahasa Indonesia memiliki *prefixes, suffixes, infexes* dan *confixes* yang membuat suatu kata dasar dapat berubah manjadi banyak bentuk dan akibatnya membuat pencarian kata dasar menjadi sulit [3]. Begitu juga dengan Bahasa Bali yang memiliki *prefixes, suffixes, infexes dan confixes* seperti Bahasa Indonesia. Berikut adalah arti dan contoh dari imbuhan dalam Bahasa Indonesia [4]:

a. *Sufiks* (Akhiran) adalah afiks yang ditambahkan pada bagian belakang kata dasar, misal "*-an, -kan,"* dan "*–i"*;
b. *Prefiks* (Awalan) adalah imbuhan yang ditambahkan pada bagian awal sebuah kata dasar atau bentuk dasar; awalan: *"per-"* adalah yang paling *produktif dalam bahasa Indonesia*
c. Konfiks (sifiks dan prefiks)afiks tunggal yang terjadi dari dua unsur yang terpisah (misal "*ke-...-an"* dalam kata "*kemerdekaan"*)

Algoritma Stemming atau *tool stemmer* untuk Bahasa Indonesia sudah banyak dikembangkan diantaranya: Nazief dan Adriani dari Universitas Indonesia pada tahun 1996, Vega dari Univertas nasional singapura tahun 2001, Arifin dan setiono dari Institut teknologi sepuluh November 2002, *Porter Stemmer for Bahasa Indonesia* dikembangkan oleh Fadillah Z. Tala pada tahun 2003 [4].

3) *Stemming* Bahasa Bali

Morfologi Bahasa Bali hampir sama dengan Bahasa Indonesia. Dimana Bahasa Bali memiliki *prefixes, suffixes, infexes dan confixes* seperti Bahasa Indonesia [8]. Hanya saja elemen pada setiap imbulan tidak semuanya sama. Berikut adalah imbulan dari Bahasa Bali:

Tabel 1. Awalan (*prefixes*)

prefixes	Replacement	contoh
ma-	Null	matinggal→tinggal, mawasta→wasta
pa-	Null	pakeweh→keweh, pamatut→patut
ka-	Null	kacarita→carita,
di-	Null	diolas→olas
sa-	Null	Sawireh→wireh
Ny-	S	Nyampat→sampat
m	p	Mancing→pancing
N	t	Nulis→tulis
ng		Ngereh→Rereh

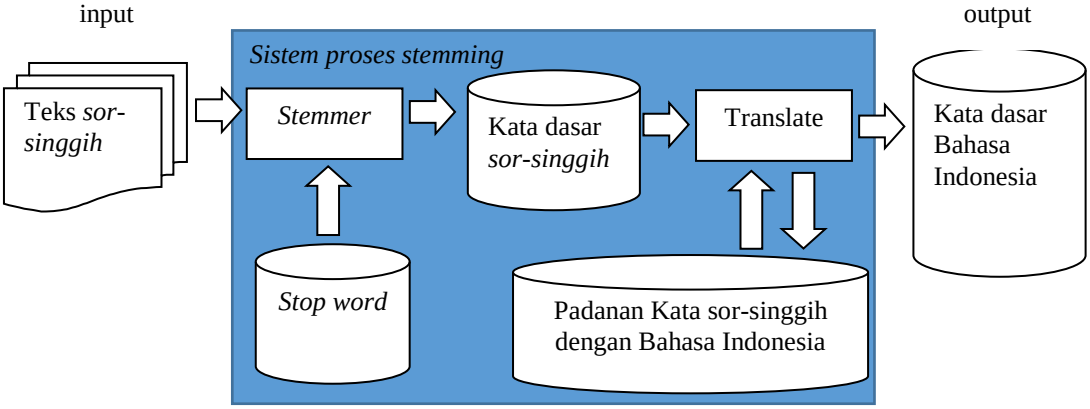
Tabel 2. Akhiran (*suffixes*)

Suffixes	Replacement	Contoh
-e	Null	Sekare→sekar
-ne	Null	Talagane→telaga, adinne→adin
-an	Null	Buungan→buung
-ang	Null	Kenehang→keneh, tuutang→tuut
-n	Null	Adin→adi
-in	Null	Sumanangsayain→sumanangsaya
-ing	Null	Angganing→anggan

Tabel 3. Konfiks (*confixes*)

Confixes	*Replacement*	Contoh
Ka-an	Null	Karahayuan →rahayu
Pa-an	Null	Palekadan →lekad

Proses *stemming* Bahasa Bali *sor-singgih* mengadopsi cara kerja dari algoritma porter *stemmer* Bahasa Indonesia. Kata dasar dari hasil *stemming* kemudian di*translate* kebahasa Indonesia. *Translate* kata dasar dari Bahasa Bali ke Bahasa Indonesia agar tingkatan kata dalam Bahasa bali dapat dirubah menjadi satu tingkat. Berikut adalah arsitektur sistem *stemmer* Bahasa Bali *sor-singgih*:



Gambar 1. Arsitektur Sistem

Proses *stemming* sor-singgih Bahasa Bali yang dirancang dapat dibagi dalam dua proses penting yaitu proses stemming dan proses *translate*. Proses *stemming* untuk menpadatkan kata dasar Bahasa Bali dan proses *translate* untuk mencari padanan kata dasar *sor-singgih* kedalam Bahasa Indonesia. Berikut adalah Penjelasan arsitektur algoritma *stemmer* Bahasa Bali.

1. Teks sor-singgih dimasukan kedalam sistem *stemming* untuk dilakukan pemotongan imbuhan (*prefixs*), mengapus kata yang kurang bermakna (*stop word*). Hasil dari proses *stemmer* berupa kata dasar dari masing – masing tingkat Bahasa / *sor-singgih*. Kata dasar sor-singgih disimpan dalam tabel kata dasar.
2. *Translate* Hasil *stemming* Bahasa bali sor-singgih kedalam Bahasa Indonesia. Proses ini menggunakan database padanan kata dasar Bahasa bali sor-singgih dengan Bahasa Indonesia. Proses ini untuk mengurangi tingkatan kata dalam Bahasa Bali.

Proses *stemmer* menggunakan cara kerja seperti algoritma porter yaitu dengan memotong imbuhan (*prefixs)*. Namun imbuhan yang digunakan hanya *prefixes* dan *suffixes.* Bahasa Bali tidak memiliki partikel baku yang secara langsung menempel pada kata dasarnya [8], jadi pada penelitian ini tidak ada proses hapus *partikel* seperti pada algoritma poter. Begitu juga dengan *possessive pronoun*,

Bahasa Bali untuk *prossesive pronoun* tidak secara langsung menempel pada kata kerja. Maka pada penelitian ini hanya mengapus awalan (*prefixs*) dan akhiran (*suffixs*).

Cara kerja algoritma *stemmer* sebagai berikut:

1. Hapus Awalan (*prefixs*).
2. Hapus akhiran (*Suffixs*) dari kata yang dimasukan
3. Simpan kata dasar pada database
4. Kata dasar sor-singgih hasil *stemming* kemudian dicarikan padanannya dengan Bahasa Indonesia agar tingkatan sor-singgih bahasa Bali di jadikan satu tingkat. Proses pencarian padanan kata menggunakan daftar padanan kata dasar.

4) Padanan kata dasar

Daftar kata yang dimaksud adalah sebuah tabel database yang menyimpan sekumpulan kata kerja dasar bahasa Bali dengan padanannya dengna kata kerja dasar Bahasa Indonesia. Padanan kata yang dicari yaitu padanan kata *sor-singgih* Bahasa Bali dengan kata dasar bahasa Indonesia. Dalam bahasa Bali setiap kata kerja pada setiap level Bahasa Bali / *sor-singgih* diucapkan atau ditulis beda namun memiliki arti yang sama. Berikut adalah beberapa contoh daftar kata (*list of word*) *sor-singgih* yang akan dicarikan padanannya dengan kata dasar Bahasa Indonesia:

Tabel 4. Padanan kata dasar

Kata dasar Bahasa Indonesia	**Kata dasar bahsa Bali**		**Kata dasar Bahasa Indonesia**	**Kata dasar bahsa Bali**
datang	rauh		dengar	mireng
	teka			dingeh
	mai		lupa	lali
	mriki			engsap
tidur	sirep		jalan	jalan
	pules			margi
	sare		cepat	gelis
	kolem			enggal
makan	amah		kecil	cenik
	ajeng			cerik
	daar			alit
	rayun		besar	gede
mati	padem			ageng
	mati			agung
	seda		anak	panak
	lebar			oka
	bangka			putra
pulang	mulih		bawa	aba
	mantuk			makta
bicara	ngandika		dapat	polih
	omong			maan
	raos		beli	numbas
	munyi			meli
	sabda			

2.2 Metode penelitian

Metode penelitian yang digunakan yaitu sebagai berikut:

1. Pengumpulan dokumen Bahasa bali
2. Analisis morfologi Bahasa bali
3. Merancang design sistem
4. Membangun *stemmer* Bahasa Bali sor-singgih sebagai uji coba.

3. Hasil dan Pembahasan

Dari hasil pengujian menggunakan kata-kata Bahasa Bali sor-singgih algoritma yang dibuat sudah mampu mencari kata dasar yang memiliki imbulan pada awalan (*prefixes*), dan akhiran (*suffixes).* Pada pengujian ini jumlah kata sor-singgih yang digunakan sejumlah 357 kata pada satu dokumen Bahasa Bali. Dari hasil pengujian 85% kata di*stemming* dengan benar. Hasil dari *stemming* yang berupa kata dasar kemudian di*translate* kebahasa Indonesia, pada proses ini jumlah kata untuk melakukan *translate* kurang banyak sehingga hanya sekitar 50% kata yang dapat di*translate* ke Bahasa Indonesia.

4. Simpulan

Berikut adalah simpulan yang dapat diambil dari penelitian ini:

1. Algoritma yang dibangun membutuhkan kata dasar lebih lebih lengkap untuk mendapatkan hasil *stemming* lebih tepat.
2. Sistem yang dibangun hanya mampu mencari kata dasar yang berisi awalan dan (atau) akhiran, sehingga belum bisa untuk kata yang berisi sisipan.
3. Dari hasil pengujian 85% kata di*stemming* dengan benar. Hasil dari *stemming* yang berupa kata dasar kemudian di*translate* kebahasa Indonesia, pada proses ini jumlah kata untuk melakukan *translate* kurang banyak sehingga hanya sekitar 50% kata yang dapat di*translate* ke Bahasa Indonesia.

Daftar Pustaka

[1] I wayan simpen. 2008. Afiksasi Bahasa Bali: sebuah kajian morfologi generative. SK Akreditasi Nomor: 007/BAN PT/Ak-V/S2/VIII/2006, Vol 15, No.29

[2] Gede Widnyana putra, sudarma made, satya kumara. (2016). Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naïve Bayes Classifier. Teknologi Elektro, Vol. 15, No.2.

[3] Asian, J., Williams, H. E., Tahaghoghi, S.M.M.,2005, *Stemming* Indonesian, Australian Computer Society Inc., Australia.

[4] Fadillah Z. Tala, 2002, "*A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*", Netherland, Universiteit van Amsterdam,

[5] MIRNA ADRIANI, Jelita Asian, Bobby Nazief, Tahagoghi, Hugh E. Williams. December-2007."*Stemming* Indonesia : A Confix-stripping approach", ACM Transactions on Asian Language Information Processing, Vol. 6, No. 4, Article 13.

[6] Ledy Agusta: 2009 Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia.

[7] Udayana Universitas Fakultas Sastra, 1977. Morfologi Bahasa Bali. Denpasar. Halaman 44-120

[8] I made Denes, ketut reoni, made pasmidi, I wayan jendra, bagus nyoman putra.1991. Morfologi Kata Benda Bahasa Bali. Departemen Pendidikan dan Kebudayaan. Jakarta