
ALGORITMA STEMMING DALAM BAHASA BALI MENGGUNAKAN PENDEKATAN N-GRAM

I Putu Satwika^{1*}, Helmy Syahk Alam²

¹Informatika, STMIK Primakara

²Sistem Informasi Akuntansi, STMIK Primakara

Email: satwika@primakara.ac.id

Abstrak: Bahasa daerah merupakan salah satu warisan bangsa yang harus tetap dilestarikan. Menurut hasil survey yang dilakukan oleh UNESCO, 11% bahasa yang ada di dunia terdapat di Indonesia. Data yang diperoleh menunjukkan bahwa 9 bahasa daerah di Papua telah punah dan terdapat lebih dari 50 bahasa daerah terancam punah. Bahasa Bali merupakan salah satu bahasa daerah yang ada di Indonesia dan tentunya memiliki ancaman terhadap kepunahan. Penelitian ini bertujuan untuk membuat suatu algoritma stemming untuk dapat memperoleh kata dasar dari suatu kata berimbuhan dalam bahasa Bali. Stemming merupakan salah satu algoritma yang sangat penting dibutuhkan dalam metode-metode perolehan informasi dan text-mining. Algoritma yang dikembangkan merupakan algoritma yang mengawali penelitian-penelitian dalam bidang text-mining dan information retrieval dalam bahasa Bali. Dari hasil penelitian yang dilakukan diketahui bahwa pendekatan dengan menggunakan teknik n-gram analisis menghasilkan akurasi yang kurang memuaskan. Dalam penelitian ini digunakan metode tambahan dengan menambahkan kamus pada algoritma yang dirancang dengan mencocokkan kata dasar dengan data yang ada pada kamus. Hasil dari penelitian menunjukkan bahwa metode tersebut mampu meningkatkan akurasi dari 54% menjadi 80% bergantung pada jumlah kata dasar yang ada pada kamus.

Kata Kunci: *stemming, Bahasa Bali, algoritma, information retrieval.*

Abstract: *Regional languages are one of the nation's heritage that must be preserved. According to the results of a survey conducted by UNESCO, 11% of the world's languages are found in Indonesia. The data obtained shows that 9 regional languages in Papua have become extinct and there are more than 50 regional languages that are threatened with extinction. Balinese is one of the regional languages in Indonesia and certainly has a threat of extinction. This study aims to create a stemming algorithm to be able to obtain basic words from affixed words in Balinese. Stemming is one of the most important algorithms needed in information-gathering and text-mining methods. The developed algorithm is an algorithm that initiates research in the fields of text-mining and information retrieval in Balinese. From the results of the research conducted, it is known that the approach using the n-gram analysis technique produces unsatisfactory accuracy. In this study an additional method was used by adding a dictionary to the algorithm designed by matching the basic words with the data in the dictionary. The results of the study show that this method can increase accuracy from 54% to 80% depending on the number of basic words in the dictionary*

Keywords: *stemming, Balinese language, algorithm, information retrieval.*

1. PENDAHULUAN

Indonesia merupakan salah satu negara dengan suku bangsa yang sangat banyak. Hal ini tentunya membuat Indonesia memiliki beragam bahasa daerah yang tersebar dari Sabang sampai Merauke. Bahasa daerah merupakan salah satu warisan budaya dari Bangsa Indonesia. Menurut UNESCO terdapat 6000 bahasa berbeda yang dituturkan di seluruh dunia dan Indonesia memiliki 700 bahasa daerah yang berarti 11% bahasa di dunia terdapat di Indonesia. Namun demikian, setengah dari bahasa-bahasa tersebut akan

punah pada abad ini. Hal ini tentunya merupakan berita yang tidak baik. Hal ini terutama terjadi di kota-kota besar yang masyarakatnya lebih heterogen sehingga membutuhkan penggunaan bahasa yang sama untuk berkomunikasi.

Apabila hal ini terus berlanjut, budaya berbahasa daerah akan semakin pudar. Hal ini akan sangat berbaya apabila terus dibiarkan yang dapat menyebabkan punahnya suatu bahasa daerah. Bahkan menurut berita harian, lebih dari 50 bahasa daerah terancam punah karena jumlah penutur yang telah lanjut usia

dan sangat sedikit anak muda yang menuturkannya. Bahkan sembilan bahasa daerah papua telah punah karena tidak ada yang menuturkannya. Berita ini juga didukung oleh dikti yang memasang berita tentang bahasa daerah yang terancam punah.

Menjaga bahasa daerah agar tidak punah merupakan tugas dan tanggungjawab kita semua terlebih untuk para putra-putri daerah yang sebenarnya bahasa ibu mereka. Bahasa daerah ini merupakan bahasa yang telah dipergunakan sejak ratusan tahun yang lalu dan telah diciptakan tidak dengan waktu yang singkat. Menjaga bahasa daerah tetap lestari memiliki peran penting untuk banyak hal seperti untuk menunjukkan identitas suku bangsa, sebagai alat perekat rasa persaudaraan di perantauan, alat ungkap kebudayaan, bagian dari mozaik kebudayaan Indonesia dan dunia, dan jembatan antargenerasi karena dengan berbahasa daerah berarti kita telah menggunakan bahasa orang tua atau leluhur.

Bahasa Bali merupakan salah satu bahasa daerah yang dituturkan oleh orang-orang suku Bali sejak lama. Bahasa ini sudah mulai ditinggalkan karena tergerus dengan kebudayaan luar. Hal ini dikarenakan salah satu dampak yang diberikan oleh berkembang pesatnya sektor pariwisata di Bali yang menjadikan masyarakat Bali mempelajari bahasa-bahasa asing dan meninggalkan bahasa daerahnya. Selain itu turis-turis lokal yang berbahasa nasional juga memberikan dampak kepada masyarakat untuk lebih mempelajari bahasa nasional daripada bahasa daerahnya.

Hal ini apabila terus dibiarkan tentunya akan menyebabkan kepunahan bahasa Bali yang menyebabkan kebudayaan daerah Bali yang terkenal hingga mancanegara akan kehilangan jati diri. Berbagai informasi tentang keberadaan kebudayaan akan sulit diungkap karena tidak ada lagi yang mampu untuk mengungkapkan dan mengerti hal tersebut. Dokumen-dokumen dalam bahasa Bali akan hanya menjadi barang di museum karena tidak ada yang mampu untuk mempelajari dan mengerti maksud dari tulisan dan bahasa yang digunakan.

Perkembangan ilmu dan teknologi diharapkan mampu untuk tetap selaras

dengan kebudayaan dan cara-cara tradisional agar tidak menggeser nilai-nilai yang telah dipupuk sejak nenek moyang. Kemanjauan teknologi seharusnya mampu memberikan kontribusi untuk menjaga dan melestarikan kebudayaan yang telah ada. Ilmu dan teknologi yang telah ada harus dapat dimanfaatkan untuk dapat menjaga kelestarian budaya.

Penelitian ini merupakan penelitian pendahu-luan yang nantinya dapat dilanjutkan untuk proses information retrieval dan text mining terhadap dokumen-dokumen yang berbahasa Bali. Pada penelitian ini akan dikembangkan metode stemming yaitu metode untuk dapat mencari kata dasar dari kata setelah diberikan imbuhan. Hal ini dikarenakan metode ini merupakan metode praproses yang sangat penting untuk dapat dipergunakan dalam berbagai metode information retrieval dan text mining. Penelitian ini sangat penting dilakukan karena belum dikembangkannya metode stemming untuk bahasa Bali.

Penelitian untuk mengembangkan metode stemming khusus untuk bahasa Bali perlu dilakukan karena metode stemming untuk bahasa lain belum tentu dapat dipergunakan pada bahasa Bali. Hal ini dapat ditunjukkan dengan metode stemming yang telah baik untuk bahasa Inggris ternyata tidak dapat digunakan untuk stemming bahasa Indonesia. Salah satu alasan mengapa tidak dapat dilakukan adalah struktur dan morfologi bahasa yang berbeda utamanya setelah suatu kata diberikan suatu imbuhan.

Dengan dikembangkannya metode stemming pada penelitian ini, diharapkan dapat mening-katkan penelitian-penelitian lain dalam bidang pe-rolehan informasi dan text mining khususnya dokumen-dokumen yang menggunakan bahasa Bali. Dengan penerapan ilmu komputer dan sistem informasi diharapkan segala dokumen dan informasi yang menggunakan bahasa Bali dapat tercatat dengan baik beserta pengertian yang terkandung dalam informasi tersebut.

Adapun luaran dari penelitian ini diharapkan mampu dipublikasikan pada terbitan berkala ilmiah internasional sehingga dapat dipelajari secara luas. Dengan mengembangkan metode pada penelitian ini

diharapkan menjadi salah satu cara kontribusi ilmu pengetahuan dan teknologi untuk dapat berperan aktif melestarikan bahasa daerah.

2. TINJAUAN PUSTAKA

Beberapa penelitian terdahulu yang penulis jadikan acuan antara lain penelitian dari Rinci Kembang Hapsari dan Yunus Juli Santoso tahun 2015 yang berjudul "Stemming Artikel Berbahasa Indonesia Dengan Pendekatan Confix-Stripping". Hasil dari penelitian ini adalah mencari kata dasar dari kata imbuhan teks berbahasa Indonesia dalam sebuah dokumen menggunakan pendekatan Confix-Stripping.

Penelitian dari Anita Guterres, Gunawan, Joan Santoso tahun 2019 yang berjudul "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based". Hasil dari penelitian ini adalah algoritma stemming bahasa Tetun menghasilkan akurasi sebesar 90.52% menggunakan pendekatan Rule Based.

Penelitian dari Rahardyan Bisma Setya Putra, Ema Utami, dan Suwanto Raharjo tahun 2018 yang berjudul "Optimalisasi Stemming Kata Berimbuhan Tidak Baku Pada Bahasa Indonesia Dengan Levenshtein Distance". Penelitian ini berfokus pada modifikasi algoritma Non-formal Affix untuk meningkatkan akurasi dalam stemming kata berimbuhan tidak baku. Hasil dari penelitian ini menunjukkan bahwa hasil modifikasi dengan Levenshtein Distance memiliki tingkat akurasi 96.6%, sedangkan algoritma Non-formal Affix memiliki akurasi 73.3% pada saat stemming 60 kata berimbuhan tidak baku. Sehingga dapat disimpulkan bahwa pendekatan dengan Levenshtein Distance dapat meningkatkan akurasi algoritma Non-formal Affix dalam melakukan stemming kata berimbuhan tidak baku.

Beberapa penelitian di atas penulis jadikan acuan untuk membuat algoritma stemming dalam bahasa Bali untuk dapat memperoleh kata dasar dari suatu kata berimbuhan dalam bahasa Bali menggunakan pendekatan n-gram.

2.1. Stemming

Yang dimaksud dengan stemming adalah salah satu cara yang digunakan untuk

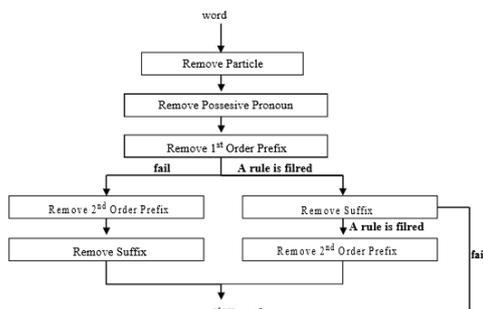
meningkatkan performa Information Retrieval dengan cara mentransformasi kata-kata dalam sebuah dokumen teks ke kata dasarnya. Algoritma Stemming untuk bahasa yang satu berbeda dengan algoritma stemming untuk bahasa lainnya. Sebagai contoh Bahasa Inggris memiliki morfologi yang berbeda dengan Bahasa Indonesia sehingga algoritma stemming untuk kedua bahasa tersebut juga berbeda. Proses stemming pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan akar kata (root word) dari sebuah kata. Beberapa algoritma stemming Bahasa Indonesia telah dikembangkan sebelumnya. Penggunaan algoritma stemming yang sesuai mempengaruhi performa sistem Information Retrieval.

Stemming merupakan proses mengembalikan berbagai macam bentuk kata ke dalam sebuah representasi yang sama. Misalnya kata "berlari" dan "dilarikan" akan diubah menjadi sebuah kata yang sama yaitu "lari". Metode stemming memerlukan masukan berupa kata yang terdapat dalam suatu dokumen, dengan menghasilkan keluaran berupa root word. Mencari kata dasar (akar kata) dari suatu kata yang berimbuhan di dalam bahasa Indonesia merupakan pekerjaan yang kompleks. Berbeda dengan bahasa Inggris yang hanya memiliki imbuhan berupa suffixes (akhiran), kata-kata dalam bahasa Indonesia bisa memiliki imbuhan yang terdiri dari prefixes (awalan), suffixes (akhiran), infixes (sisipan) dan confixes (kombinasi dari awalan, akhiran dan sisipan). Penggunaan imbuhan di dalam Bahasa Indonesia bisa memakai lebih dari satu prefix dan sufiks dalam satu kata. Sebuah kata dengan awalan bisa memiliki arti yang berbeda dari kata dasarnya. Hal ini disebut derivation word. Selain itu sebuah awalan juga dapat mengubah huruf awal suatu akar kata yang biasa disebut peluluhan, sehingga dibutuhkan pengkodean ulang (recode) untuk menemukan kembali akar kata berdasarkan penggunaan awalan.

Menurut penelitian yang dilakukan oleh Ledy Agusta, terdapat dua algoritma stemming yang sesuai dengan morfologi untuk Bahasa Indonesia yaitu Algoritma Porter dan

Algoritma Nazief & Adriani. Dari hasil penelitian tersebut didapatkan kesimpulan bahwa:

- a. Proses stemming dokumen teks berbahasa Indonesia menggunakan Algoritma Porter membutuhkan waktu yang lebih singkat dibandingkan dengan stemming menggunakan Algoritma Nazief & Adriani.
- b. Proses stemming dokumen teks berbahasa Indonesia menggunakan Algoritma Porter memiliki prosentase keakuratan (presisi) lebih kecil dibandingkan dengan stemming menggunakan Algoritma Nazief & Adriani.
- c. Pada proses stemming menggunakan Algoritma Nazief & Adriani, kamus yang digunakan sangat mempengaruhi hasil stemming. Semakin lengkap kamus yang digunakan maka semakin akurat pula hasil stemming.



Gambar 2.1 Desain algoritma Porter

Pada gambar 2.1 dapat diperhatikan skema dari algoritma porter untuk melakukan proses pencarian kata dasar suatu kata berimbuhan. Adapun langkah-langkah algoritma ini adalah sebagai berikut:

1. Hapus Particle.
2. Hapus Possesive Pronoun.
3. Hapus awalan pertama. Jika tidak ada lanjutkan ke langkah 4a, jika ada cari maka lanjutkan ke langkah 4b.
4. Hapus awalan kedua, lanjutkan ke langkah 5a.
5. Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai root word. Jika ditemukan maka lanjutkan ke langkah 5b.
6. Hapus akhiran. Kemudian kata akhir diasumsikan sebagai root word.

7. Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai root word.

2.2. Imbuhan Bahasa Bali

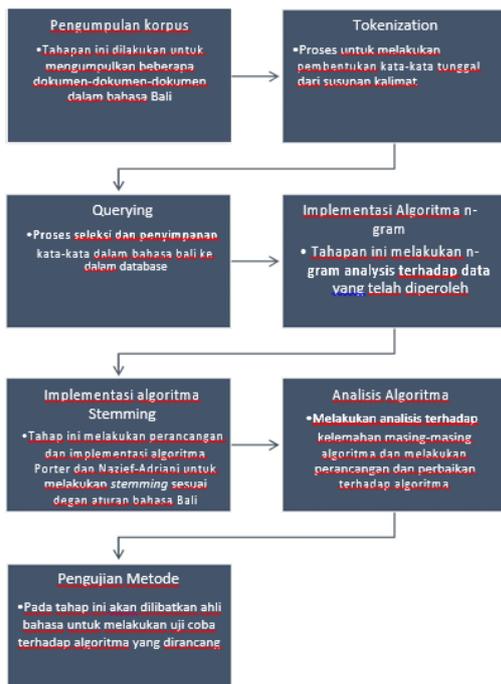
Dalam bahasa Bali dikenal terdapat beberapa imbuhan yang hampir mirip dengan bahasa Indonesia yaitu pengater (awalan), seselan (sisipan), pengiring (akhiran), pengater muah pengiring (awalan dan akhiran). Dalam penambahan imbuhan juga terjadi peluruhan seperti yang terjadi dalam bahasa Inggris dan bahasa Indonesia. Adapun kata yang telah memperoleh imbuhan dalam bahasa Bali disebut dengan krana tiron

3. METODE

Pada penelitan ini dilakukan beberapa langkah yang harus dilakukan sebelum memperoleh hasil stemming dari suatu kata. Adapun proses tersebut dapat diperhatikan sesuai dengan gambar 1.

Adapun teknik pengumpulan data yang dilakukan adalah dengan mengumpulkan berbagai dokumen yang menggunakan bahasa Bali untuk dijadikan sebagai Korpus. Korpus ini yang akan digunakan sebagai bahan analisis terhadap kata-kata yang ada dalam bahasa Bali. Korpus dapat diperoleh dari berbagai media baik elektronik maupun cetak yang tentunya semua media tersebut harus dikonversi terlebih dahulu kedalam bentuk digital agar dapat diolah menggunakan computer.

Pada penelitian ini dilakukan dengan menggunakan pendekatan eksperimen. Eksperimen dilakukan dengan membandingkan akurasi dari proses stemming kata berimbuhan menjadi bentuk kata dasarnya. Semakin banyak kata berimbuhan yang dapat diubah dengan benar ke dalam bentuk kata dasarnya, maka kinerja metode yang dikembangkan juga semakin baik. Untuk proses validasi terhadap stemming ini, penelitian ini juga melibatkan beberapa pakar yang berkecimpung dalam kesusastraan Bali. Para pakar ini akan melakukan validasi terhadap hasil stemming dari algoritma yang dikembangkan.



Gambar 3.1. Alur Algoritma Stemming Dengan Pendekatan Analisis N-gram

4. HASIL DAN PEMBAHASAN

Proses input data terhadap kata-kata berimbuhan dan kata-kata dasar dalam bahasa Bali dilakukan secara manual ke dalam basisdata. Pada awal penelitian digunakan sejumlah 523 kata berimbuhan berikut dengan kata dasarnya. Untuk melakukan pengujian maka dilakukan implementasi terhadap analisis n-gram. Dari hasil analisis, maka selanjutnya dibuatlah algoritma stemming untuk mencari kata dasar dalam bahasa Bali. Berikut adalah algoritma yang diimplementasikan pada awal penelitian ini.

- A. Proses penghilangan seselan
 1. IF kata IN “kamus kata seselan” THEN REMOVE *-in-, -um-, -el-, -er-*.
- B. Proses penghilangan awalan
 2. IF kata diawali oleh awalan kategori_1 (*a-, pra-, pari-, pati-, maka-, saka-, kuma-, ka-, sa-*) THEN:
 - a. IF awalan *a-*
 - i. REMOVE *a-*
 - ii. IF kata_dasar “kruna wilangan” and kata dasar > 3 grams THEN kata_dasar TRUE
 - iii. ELSE kata_dasar FALSE

- b. ELSE
 - i. REMOVE prefix-
 - ii. IF kata_dasar > 3 grams THEN kata_dasar TRUE
 - iii. ELSE kata_dasar FALSE
3. IF kata diawali oleh awalan kategori_2 (*pa-, pi-*) AND kata_dasar > 3 grams THEN
 - c. IF awal kata_dasar 2 KONSONAN THEN REMOVE *pa-/pi- + 1 KONSONAN*
 - d. ELSE REMOVE *pa-/pi-*
4. IF kata diawali oleh awalan kategori_3 (*ma-*) THEN
 - e. IF kata_dasar > 3 grams THEN IF huruf awal kata_dasar VOKAL THEN REMOVE *m-* ELSE IF huruf awal kata_dasar KONSONAN THEN REMOVE *ma-*
 - f. ELSE kata_dasar FALSE
5. IF kata diawali oleh awalan kategori_4 Anusvara (*ng-, ny-, nge-, nga-*) and kata_dasar > 3 grams THEN
 - g. IF awalan *ng-* THEN
 - i. IF huruf awal kata_dasar VOKAL THEN REMOVE *ng-* AND ADD *g-/k-* di awal huruf.
 - h. IF awalan *ny-* THEN REMOVE *ny-* AND ADD *c-/j-/s-* di awal huruf
 - i. IF awalan *nge-/nga-* THEN REMOVE *nge-/nga-*

- C. Proses penghilangan akhiran
 6. IF kata diakhiri oleh akhiran kategori_1 (*-ing, -ne*) AND kata_dasar > 3 grams THEN REMOVE *-ing/-ne*
 7. IF kata diakhiri oleh akhiran kategori_2 (*-e, -a, -n*) THEN kata_dasar > 4 grams THEN REMOVE *-e/-a/-n*
 8. IF kata diakhiri oleh akhiran kategori_3 (*-ang, -in, -an*) and kata_dasar >= 4 grams THEN
 - a. IF akhir kata_dasar 2 VOKAL THEN REMOVE *-ang/-in/-an* ELSE IF akhir kata_dasar *-n* THEN REMOVE *-nang/-nin/-nan*

Dari hasil implementasi algoritma di atas dapat diketahui bahwa sistem yang dibuat belum dapat menghasilkan akurasi yang baik. Hal ini dikarenakan oleh kompleksitas dari penyusunan kata berimbuhan dalam bahasa

Bali. Selanjutnya adalah penjelasan mengenai permasalahan yang dihadapi dalam implementasi algoritma versi pertama ini yang mengakibatkan nilai akurasi proses stemming menggunakan algoritma yang dirancang tidak berhasil dengan baik.

Dalam melakukan penelitian diketahui bahwa kata yang memperoleh imbuhan seselan (sisipan) dalam Bahasa Bali jumlahnya sangat terbatas. Maka dari itu implementasi digunakan dengan teknik hardcode pada bahasa pemrograman sehingga akan lebih mudah dilakukan dan mengurangi kompleksitas algoritma ini. Hal ini dapat terjadi karena tidak diperlukan penambahan proses loop dan condition pada algoritma yang dibuat, sehingga proses pencarian kata dasar menjadi lebih efisien.

Di dalam bahasa Bali terdapat banyak imbuhan yang apabila melekat pada kata dasar tertentu maka terjadi proses peluruhan. Hal ini juga menjadi salah satu hal yang menyebabkan proses pencarian bentuk kata dasar dari kalimat berimbuhan dalam bahasa Bali menjadi sangat rumit. Berikut adalah contoh dari peluruhan tersebut:

1. N- + kancing _ ngancing, dimana huruf k- menjadi ng-
2. N- + tunas _ nunas, dimana huruf t- menjadi n-
3. Ma- + embon _ membon, dimana terjadi penggabungan huruf vokal mengikuti kata dasarnya.

Selain itu, dalam bahasa Bali juga terdapat penambahan huruf jika kata dasar diberikan suatu imbuhan tertentu. Sebagai contoh adalah kata-kata berikut ini:

1. Gede + -ang _ gedenang, terdapat penambahan huruf -n-
2. Beli + -in _ belinin, terdapat penambahan huruf -n-

Dalam bahasa Bali juga terdapat kata dasar yang bila diberikan imbuhan yang sama dapat berubah menjadi dua bentuk kata yang berbeda namun memiliki makna yang sama. Sebagai contoh kata tersebut adalah sebagai berikut:

1. Gede + -ang _ gedeang/gedenang
2. Sauta + -ang _ satuang/satuayang

3. Bucu + -in _ bucuin/bucunin
4. Sari + -in _ sarinin/sariin.
5. Malu + -an _ maluan/malunan
6. Ciri + -an _ ceciren

Dari hasil penelitian diketahui bahwa imbuhan anusuara (n-) merupakan imbuhan yang aturan pembentukannya paling rumit dan kompleks bila dibandingkan dengan imbuhan lainnya. Hal inilah yang menyebabkan proses stemming menggunakan metode yang telah dikembangkan gagal memperoleh akurasi yang tinggi. Hal ini dikarenakan sistem tidak dapat mengetahui kata dasar dari suatu kata berimbuhan akibat kompleksitas pembentukan kata yang memperoleh imbuhan anusuara ini. Sebagai contoh aturan pembentukan kata yang memperoleh imbuhan anusuara (n-) pada suatu kata dasar yang diawali oleh huruf k dan g, maka awalan anusuara (n-) akan berubah menjadi ng-

. Misalnya kata "kancing" menjadi "ngancing" dan kata "gebeg" menjadi "ngebeg". Algoritma yang telah dibuat akan mengalami kebingungan untuk melakukan proses stemming karena kata berimbuhan "ngebeg" mungkin menjadi kata "kebeg" (tidak memiliki arti) atau "ge beg" (menggosok). Begitu juga kata "ngancing" yang seharusnya menjadi "kancing" dan "g ancing" (tidak memiliki arti). Kedua pembentukan kata ini tidak memiliki ciri-ciri khusus yang menyebabkan proses stemming terhadap kata-kata tersebut menjadi rumit dan sulit.

Selain kata-kata bentuk dasar, dalam bahasa Bali juga dikenal bentuk-bentuk kata lain yang dapat berimbuhan seperti kruna satma (kata majemuk) dan kruna dwi lingga (kata ulang). Dalam penelitian ini tidak dibahas proses pembentukan kedua jenis kata ini. Sistem yang dibuat dalam penelitian ini tidak melakukan proses stemming terhadap kedua jenis kata tersebut. Hal ini untuk membuat algoritma yang dibuat agar lebih sederhana mengingat waktu yang cukup terbatas.

Dari hal-hal di atas dapat diketahui bahwa proses pencarian kata dasar dalam bahasa Bali sangatlah rumit dan kompleks. Tim peneliti mencoba untuk melakukan beberapa metode untuk mampu meningkatkan hasil akurasi proses stemming, namun tidak diperoleh hasil yang cukup signifikan. Akhirnya tim peneliti mencoba untuk melakukan studi literatur terhadap

proses stemming bahasa lainnya untuk mengetahui apakah terdapat penelitian yang memiliki kendala yang sama dengan penelitian yang dilakukan.

Dari hasil diskusi dan studi literatur yang cukup mendalam, maka tim sepakat untuk melakukan penambahan proses agar permasalahan tersebut dapat diatasi. Untuk mengatasi hal tersebut, maka tim peneliti mencoba untuk berpikir selayaknya manusia berpikir. Manusia dapat mengetahui suatu kata ada atau tidak jika (memiliki arti atau tidak) apabila pernah mendengar dan menghafalkan kata tersebut. Maka dari itu, pada penelitian ini kami menambahkan sebuah kamus kata dasar. Jika suatu kata setelah proses stemming diperoleh lebih dari satu kata yang mungkin, maka kata-kata tersebut selanjutnya dicocokkan pada data kata dasar yang ada di database. Jika terdapat kata yang cocok, maka kata tersebut merupakan kata yang tepat sebagai kata dasarnya.

Pengujian terhadap algoritma dengan kamus kata dasar diperoleh hasil bahwa algoritma yang dikembangkan mampu meningkatkan hasil akurasi perubahan kata berimbuhan ke kata dasar. Selain itu, dapat diketahui juga bahwa semakin lengkap kamus yang dibuat, maka akurasi yang diperoleh semakin baik pula. Berikut adalah rancangan algoritma setelah berhasil di optimasi:

- A. Proses penghilangan seselan
 1. IF kata IN "kamus kata seselan" THEN REMOVE *-in-, -um-, -el-, -er-*.
- B. Proses penghilangan awalan
 2. IF kata diawali oleh awalan kategori_1 (*a-, pra-, pari-, pati-, maka-, saka-, kuma-, ka-, sa-*) THEN:
 - a. IF awalan *a-*
 - i. REMOVE *a-*
 - ii. IF kata_dasar "kruna wilangan" and kata dasar > 3grams THEN kata_dasar TRUE
 - iii. ELSE kata_dasar FALSE
 - b. ELSE
 - i. REMOVE prefix-
 - ii. IF kata_dasar > 3 grams THEN kata_dasar TRUE
 - iii. ELSE kata_dasar

FALSE

3. IF kata diawali oleh awalan kategori_2 (*pa-, pi-*) AND kata_dasar > 3 grams THEN
 - b. IF awal kata_dasar 2 KONSONAN THEN REMOVE *pa-/pi- + 1* KONSONAN
 - c. ELSE REMOVE *pa-/pi-*
4. IF kata diawali oleh awalan kategori_3 (*ma-*) AND kata_dasar > 3 grams THEN
 - a. IF huruf awal kata_dasar VOKAL THEN REMOVE *m-*
 - b. ELSE IF huruf awal kata_dasar KONSONAN THEN REMOVE *ma-*
 - c. IF kata_dasar NOT-IN "kamus" THEN kata_dasar FALSE
5. IF kata diawali oleh awalan kategori_4 Anusvara (*ng-, ny-, nge-, nga-*) and kata_dasar > 3 grams THEN
 - d. IF awalan *ng-* THEN
 - i. IF huruf awal kata_dasar VOKAL THEN REMOVE *ng-* AND ADD *g-* di awal huruf
 - ii. IF kata_dasar NOT-IN "kamus" THEN kata_dasar ADD *k-*
 - iii. IF kata_dasar NOT-IN "kamus" THEN kata_dasar FALSE
 - e. IF awalan *ny-* THEN REMOVE *ny-* AND ADD *c-* di awal huruf
 - i. IF kata_dasar NOT-IN "kamus" THEN kata_dasar ADD *j-* di awal huruf
 - ii. IF kata_dasar NOT-IN "kamus" THEN kata_dasar ADD *s-* di awal huruf
 - iii. IF kata_dasar NOT-IN "kamus" THEN kata_dasar FALSE.
 - c. IF awalan *nge-/nga-* THEN REMOVE *nge-/nga-*.

Proses penghilangan
akhiran

6. IF kata diakhiri oleh akhiran kategori_1 (-ing, -ne) AND kata_dasar > 3 grams THEN REMOVE - ing/-ne
7. IF kata diakhiri oleh akhiran kategori_2 (-e, -a, -n) THEN kata_dasar > 4 grams THEN REMOVE - e/-a/-n
8. IF kata diakhiri oleh akhiran kategori_3 (-ang, -in, -an) and kata_dasar >= 4 grams THEN
 - a. IF akhir kata_dasar 2 VOKAL THEN REMOVE - ang/-in/-an
 - b. ELSE IF akhir kata_dasar - n THEN REMOVE -nang/-nin/-nan

5. KESIMPULAN DAN SARAN

Dari hasil penelitian yang dilakukan dapat diketahui bahwa proses pembentukan kata berimbuhan dalam bahasa Bali memiliki kompleksitas yang cukup rumit. Hal ini dapat ditunjukkan dengan perubahan huruf konsonan maupun vokal baik dengan melakukan peluruhan maupun perubahan bentuk apabila melekat pada suatu imbuhan tertentu. Hal ini menyebabkan pendekatan stemming dalam bahasa Bali dengan pendekatan analisis N-Gram mengalami kelemahan dengan menghasilkan akurasi yang kurang baik. Algoritma ini selanjutnya dioptimasi dengan menambahkan "kamus" yang menyimpan bentuk-bentuk kata dasar dalam bahasa Bali. Nantinya setelah diproses melalui algoritma ini, kata dasar dicocokkan dengan data yang ada pada "kamus". Hal ini mampu meningkatkan akurasi dari algoritma ini.

Sebagai saran untuk penelitian selanjutnya, para peneliti dapat melakukan modifikasi dan optimasi terhadap algoritma yang telah dirancang dengan metode lainnya. Selain itu, pada penelitian selanjutnya dapat dilakukan dengan menggunakan pendekatan metode lain seperti statistik untuk mencari

kata dasar terhadap kata-kata berimbuhan dalam bahasa Bali.

DAFTAR REFERENSI

- [1] UNESCO, "Biodiversity and linguistic diversity," 23 April 2014. [Online]. Available: http://www.unesco.org/new/en/culture/the_mses/endangered_languages/biodiversity-and-linguistic-diversity/.
- [2] Rinci Kembang Hapsari dan Yunus Juli Santoso, "Stemming Artikel Berbahasa Indonesia Dengan Pendekatan Confix-Stripping". Prosiding Seminar Nasional Manajemen Teknologi XXII, 2015.
- [3] Anita Guterres, Gunawan, Joan Santoso, "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based", 2019.
- [4] Rahardyan Bisma Setya Putra, Ema Utami, Suwanto Raharjo, "Optimalisasi Stemming Kata Berimbuhan Tidak Baku Pada Bahasa Indonesia Dengan Levenshtein Distance". Jurnal Pengembangan IT (JPIT), Vol.03, No.02, 2018.
- [5] DPA, "Home: 140 Bahasa Daerah Di Indonesia Terancam Punah," 23 April 2014. [Online]. Available: <http://www.suarapembaruan.com/home/140-bahasa-daerah-di-indonesia-terancam-punah/50053>.
- [6] JPNN, "Berita: Ratusan Bahasa Daerah Terancam Punah," 05 September 2012. [Online]. Available: <http://www.dikti.go.id/id/2012/09/05/ratusan-bahasa-daerah-terancam-punah/>.
- [7] A. Nazief and M. Adriani, "Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia (Citations: 2)," in ACM Transactions on Asian Language Information Processing, 1996.
- [8] L. Agusta, "Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia," in Konferensi Nasional Sistem dan Informatika, Bali, 2009.
- [9] Dep. P&K, Tata Bahasa Bali: Proyek Pengembangan Bahasa dan Sastra

-
- Indonesia dan Daerah, Denpasar, 1984/1985.
- [10] I. Tinggen, Pedoman Perubahan Ejaan Bahasa Bali dengan Huruf Latin dan Huruf Bali, Denpasar, 1987.
- [11] P. D. T. I. Dinas Pengajaran Bali, Ejaan Bahasa Daerah Bali yang Disempurnakan (Huruf Latin), Denpasar, 1990.
- [12] A. HANAFI, "Encyclopedia: Metode N-Gram," Digital Library ITS, 30 April 2009. [Online]. Available: http://digilib.ittelkom.ac.id/index.php?option=com_content&view=article&id=531:metode-n-gram&catid=20:informatika&Itemid=14. [Accessed 23 April 2014].
- [13] F. RAHMAWAN, Implementasi Question Answering System pada Dokumen Bahasa Indonesia menggunakan Metode N-Gram, Bogor: Fak. MIPA, IPB, 2011.
- [14] L. Sedy Andrian Sugianto, "PEMBUATAN APLIKASI PREDICTIVE TEXT MENGGUNAKAN METODE N-GRAM-BASED," JURNAL INFRA, vol. I, no. 2, 2013.
- [15] H. Sujaini, A. Purwarianti, A. Arman and Kuspriyanto, "Extended word similarity based clustering on unsupervised PoS induction to improve English-Indonesian statistical machine translation," in Conference on Asian Spoken Language Research and Evaluation, 2013.