Modifikasi Algoritma Porter Stemmer Untuk Stemming Bahasa Sasak

Yulita Fatma Andriani[1)], Ema Utami[2)], Suwanto Raharjo[3)]

[1,2,3]Magister Teknik Informatika Universitas Amikom Yogyakarta
[1,2,3]Condong Catur, Depok, Sleman, Yogyakarta 55281, Indonesia
[1]yulita.andriani@students.amikom.ac.id, [2]ema.u@amikom.ac.id

Abstrak

Internet atau sering disebut dengan dunia maya juga berkembang dengan pesat, dunia maya akhir-akhir ini digunakan untuk sarana jejaring sosial. Tidak sedikit masyarakat menuangkan opininya pada social media. Dengan teknologi saat ini opini masyarakat sudah bisa diekstraksi menggunakan text mining di social media, dengan begitu pihak lain (pemerintah ataupun perusahaan swasta) dapat mengetahui opini masyarakat. Opini masyarakat Sasak dapat diekstraksi dengan algoritma yang sudah ada dapat dilakukan stemming tetapi ketika algoritma tersebut diaplikasikan pada bahasa yang berbeda bisa saja hasil yang diinginkan tidak sesuai, oleh karena itu penelitian ini diajukan untuk stemming bahasa Sasak. Penelitian ini menggunakan porter stemmer untuk stemming Bahasa Sasak dengan tingkat keberhasilan stemming 80%.

Kata kunci: Porter Stemmer, Stemming, Bahasa Sasak

1. PENDAHULUAN

Bahasa Sasak adalah bahasa suku Sasak. Pada umumnya, bagi masyarakat kebanyakan di Lombok, hanya dikenal dua bentuk bahasa dalam komunikasi sehari-hari, yaitu yang disebut dengan bahasa Sasak biase/jamaq atau aok-ape (ya-apa) dan Sasak alus atau tiang-enggih (saya-ya). Adapun bahasa Sasak sangat halus, yang disebut kaji-meran (saya-ya), hanya dipakai oleh para datu-raden (raja dan kaum perwangsa atau ningrat). Klasifikasi itu didasarkan pada stratifikasi sosial masyarakat Sasak sebagai bangsawan atau menak (perwangsa) dan bukan bangsawan atau non-menak (Wiliam, 2006).

Internet atau sering disebut dengan dunia maya juga berkembang dengan pesat, dunia maya akhir-akhir ini digunakan untuk sarana jejaring sosial. Tidak sedikit masyarakat menuangkan opininya pada social media. Dengan teknologi saat ini opini masyarakat sudah bisa diekstraksi menggunakan text mining di social media, dengan begitu pihak lain (pemerintah ataupun perusahaan swasta) dapat mengetahui opini masyarakat.

Untuk mendapatkan opini masyarakat tersebut, perlu dilakukan stemming pada data-data yang ada di social media. Oleh karena itu peneliti mengembangkan algoritma untuk stemming bahasa Indonesia. Bahkan terdapat penelitian bertujuan meningkatkan hasil stemming bahasa indonesia untuk bahasa gaul (slang) (Maylawati, 2018).

Untuk bahasa Bali dalam penelitian yang menggunakan algoritma Porterm stemmer (Nata, 2017) dan penelitian yang menggunakan pendekatan rule-based dan metode n-gram stemming (Subali, 2019). Dengan adanya penelitian tersebut opini masyarakat dapat diekstrak dalam berbagai bahasa.

Opini masyarakat Sasak dapat diekstraksi dengan algoritma yang sudah ada dapat dilakukan stemming tetapi ketika algoritma tersebut diaplikasikan pada bahasa yang berbeda bisa saja hasil yang diinginkan tidak sesuai, oleh karena itu penelitian ini diajukan untuk stemming bahasa Sasak.

Penelitian ini dilakukan berdasarkan banyaknya penelitian terdahulu tentang data mining dan stemming, tetapi belum ada penelitian untuk bahasa Sasak. Selain agar lebih banyak mengenai penelitian terdahulu juga menjadikan bahan perbandingan metode manakah yang memiliki evaluasi terbaik.

2. METODE PENELITIAN

Metode yang digunakan adalah dengan memodifikasi porter stemmer agar bisa digunakan untuk stemming bahasa sasak. Eksperiman yang dilakukan adalah dengan Porter Stemmer untuk mengetahui apakah algoritma ini memiliki akurasi yang cukup tinggi dalam melakukan stemming bahasa Sasak. Dilihat dari segi sifatnya penelitian ini adalah penelitian bersifat deskriptif dengan pemahaman masalah yang baik dan data yang terstruktur.

Dalam penelitian ini, pengumpulan data yang akan digunakan menggunakan beberapa langkah yang berkaitan dengan metode penelitian tersebut yaitu dengan metode observasi dan studi literatur atau kepustakaan

3. TINJAUAN PUSTAKA

Pada penelitian stemming Bahasa Bali menggunakan pendekatan metode Porter Stemmer untuk mengembangkan algoritma stemming khusus untuk Bahasa Bali dalam menanggani sor-singgih pada dukumen bahasa Bali. Dalam peneltian tersebut menjelaskan adanya 2 langkah yang dilakukan peneliti. pertama stemming, pemotongan kata untuk mendapatkan kata dasar bahasa bali. kemudian langah selanjutnya adalah menerjemahkan ke bahasa indonesia menggunakan mapping kata dasar bali-indonesia yang sudah disimpan dalam database (Nata, 2017).

Penelitian stemming pada bahasa Bali bertujuan mengembangkan metode stemmer yang meluluhkan seluruh variasi afiks pada bahasa Bali dengan mengombinasikan pendekatan rule-based dan metode N-gram stemming. Untuk kesepuluh query metode yang diusulkan memperoleh rerata akurasi stemming lebih baik 96,67% dari metode terdahulu 75%, sedangkan untuk kelima query metode n-gram stemming dapat mengenali beberapa kata berafiks diluar rules (Subali, 2019).

Satu lagi penelitian yang memodifikasi Porter Stemmer dikembangkan khusus untuk plugin error detector. Algoritma asli Porter dan algoritma modifikasi Porter mampu menganalisis dengan sempurna semua kata yang salah (100%), namun, masih ada kekurangan dalam menganalisis kata-kata yang benar. penggunaan algoritma Porter yang dimodifikasi menghasilkan hasil yang lebih baik dalam menganalisis kata-kata yang benar. Keakuratan algoritma yang dimodifikasi dalam menganalisis kata-kata yang benar adalah 96,31%, sedangkan akurasi algoritma asli Porter adalah 93,04%. Algoritma Porter yang dimodifikasi memiliki algoritma yang lebih kompleks dan awalan, postfix, dan tabel aturan sufiks yang lebih lengkap. Algoritma Porter yang dimodifikasi dirancang untuk meminimalkan kesalahan dan kekurangan algoritma asli Porter dalam proses menganalisis kata-kata yang benar (Widjaja, 2015).

Proses untuk mengekstraksi kata dasar dari kata berafiks dikenal dengan istilah stemming (Balasankar, 2016) yang bertujuan meningkatkan recall dengan mereduksi variasi kata berafiks ke dalam bentuk kata dasarnya (Patil, 2017) (Pramudita, 2018).

Dalam penjelasan lain juga disebutkan bahwa Stemming merupakan suatu metode bagian dari NLP (Putra, 2018) Stemming adalah suatu proses pelepasan afiks dari sebuah kata (Bird, 2014). Pada stemming biasanya terdapat error understemming dan overstemming. Understemming adalah ketika kata terlaly terlalu sedikit dipotong imbuhannya, sedangkan overstemming adalah ketika kata terlalu banyak dipotong imbuhan (Kaara, 2013).

Algoritma Porter stemming adalah stemmer konflasi yang diusulkan oleh Porter. Algoritma ini didasarkan pada gagasan bahwa imbuhan dalam bahasa Inggris sebagian besar terdiri dari kombinasi imbuhan yang lebih kecil dan lebih sederhana. Proses pengupasan dilakukan pada serangkaian langkah, khususnya lima langkah, yang mensimulasikan proses infektif dan turunan kata. Pada setiap langkah, imbuhan tertentu dihapus dengan aturan substitusi. Aturan substitusi diterapkan ketika seperangkat kondisi / kendala yang melekat pada aturan ini berlaku. Salah satu contoh kondisi seperti itu adalah panjang minimal (jumlah urutan konsonan-vokal) dari batang yang dihasilkan. Panjang minimum ini disebut ukuran. Kondisi sederhana lainnya pada batang dapat berupa apakah batang berakhir dengan konsonan, atau apakah batang mengandung vokal.

Ketika semua kondisi aturan tertentu terpenuhi, aturan diterapkan, yang menyebabkan penghapusan imbuhan dan kontrol bergerak ke langkah berikutnya. Jika kondisi aturan tertentu dalam langkah saat ini

tidak dapat dipenuhi, kondisi aturan berikutnya dalam langkah tersebut diuji, sampai aturan tersebut selesai atau aturan dalam langkah itu habis. Proses ini berlanjut untuk semua lima langkah (Tala, 2002).

Morfologi adalah telaah secara structural terhadap morfem-morfem beserta penyusunannya dalam rangka pembentukan kata yang banyak terpakai dalam bahasa Sasak. Morfem dalam bahasa Sasak terbagi menjadi dua, yaitu morfem bebas dan morfem terikat. (Hakim, 2016).

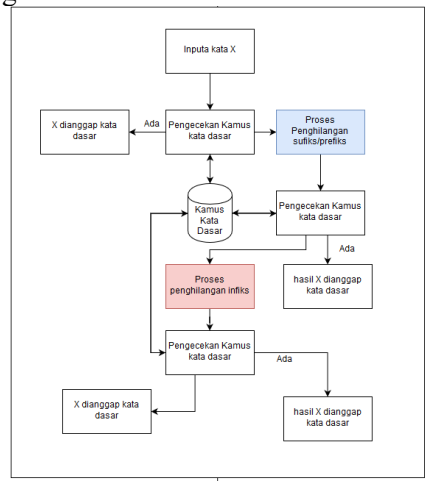
Morfem bebas dalam penggunaan bahasa Sasak ada beberapa contoh, contoh-contoh ini dapat dikatakan morfem bebas karena dapat berdiri sendiri dan dapat diucapkan tersendiri walaupun tidak diletakkan dalam hubungan kalimat. Morfem terikat adalah morfem yang tidak dapat berdiri sendiri, mengandung makna setelah dipadukan dengan morfem laun atau bentuk lain. Morfem terikat terbagi menjadi dua yaitu morfem yang terikat secara morfologis dan morfem terikat secara sintaksis. Contoh prefiks dalam bahasa sasak dapat dilihat pada tabel 1.

Tabel 1. Contoh Prefiks

Prefiks	Morfem terikat
/be-/	Begawe, bekedek, bedait, beruni, berujuk
/ber-/	Berobah, berongkos, beradat
/pe-/	Pemaling, penyopet, penginem
/peng-/	Pengkedek, pengawis
/me-/	Menyusah, memaling, meliwat
/nge-/	Ngeraos, ngengakoq, ngengais

4. HASIL DAN PEMBAHASAN

Flow untuk stemming ini dapat dilihat pada gambar 1.



Gambar 1. Flow Stemming

Terdapat beberapa kemungkinan scenario yang dijalankan:

a. Kata yang dimasukkan merupakan kata dasar. Terdapat skenario dimana kata yang dimasukan adalah kata dasar. Pada Gambar 1, sebelum menjalankan algoritma maka akan dicek dulu apakah kata yg dimasukkan sudah ada di kamus dasar kata. Jika kata tersebut ada di kamus dasar kata berarti stemming tidak akan dilakukan. contoh kata yang dimasukkan adalah lalo yang berarti pergi. kata lalo tidak memilik imbuhan sama sekali, dengan begitu lalo adalah kata dasar.
b. Kata yang dimasukkan memiliki imbuhan tanpa sisipan. Jika kata yang dimasukkan memiliki imbuhan atau bukan merupakan kata dasar maka hasil pencarian di kamus dasar kata bernilai false. Setelah hasil penghilangan prefiks/sufiks didapatkan maka akan dicek lagi menggunakan kamus kata dasar. Jika hasil stemming ada di kamus kata dasar maka proses dihentikan. Contoh menggunakan kata *tetaletan*, kata tersebut memiliki kata dasara talet yang berarti tanam dan imbuhan te-an yang mengubah artinya menjadi tanaman.
c. Kata yang dimasukkan memiliki imbuhan dengan sisipan. Jika kata yang dimasukkan memiliki sisipan, maka proses yang dilakukan sama seperti point b) tetapi hasil pencarian di kamus kata untuk kedua kalinya tentunya akan bernilai false. Karena masih bernilai false, maka akan dilanjutkan stemming. Terlepas dari ada atau tidaknya hasil dari algoritma tersebut terdapat di kamus kata dasar, point ini memberikan hasil akhir. Jika hasil stemming memang ada di kamus kata dasar maka stemming berhasil, jika tidak ada kemungkinan terdapat overstemming atau understemming tetapi hasil ini akan tetap ditampilkan. Contoh menggunakan kata *begegoloqan* yang memiliki arti tidur-tiduran, be-an merupakan simufliks (gabungan awalan dan akhiran) yang kemudian menyisakan *gegoloq* dengan sisipan -eg-. Kata dasar dari begegoloqan adalah goloq.

Dari hasil pengujian menggunakan kata-kata Bahasa Sasak algoritma yang dibuat sudah mampu mencari kata dasar yang

memiliki imbulan pada awalan (prefixes), dan akhiran (suffixes). Pada pengujian ini jumlah kata bahasa yang digunakan sejumlah 200 kata pada satu teks cerita Bahasa Sasak. Dari hasil pengujian 80% kata distemming dengan benar. Hasil dari stemming yang berupa kata dasar.

5. KESIMPULAN DAN SARAN

a. Kesimpulan

Algoritma yang dibangun membutuhkan kata dasar yang lebih lengkap untuk mendapatkan hasil stemming lebih akurat. Sistem yang dibangun hanya mampu mencari kata dasar yang berisi awalan dan/atau akhiran, sehingga belum bisa untuk kata yang berisi sisipan. Dari hasil pengujian 80% kata distemming dengan benar. Hasil dari stemming yang berupa kata dasar.

b. Saran

Saran untuk pengembangan adalah memperhatikan tingkatan bahasa Sasak kemudian ditranslate ke bahasa lain.

6. REFERENSI

Bird, S. (2014). *Natural Language Processor.* O'REILLY.

Hakim, L. (2016). *Ensiklopedia Bahasa Sasak.* Mataram: Kantor Bahasa Nusa Tenggara Barat.

Kaara, W. B. (2013). A New Stemmer to improve InformationRetrieval. *International Journal of Network Security & Its Application (IJNSA)*.

Maylawati, D. S. (2018). An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media. *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*.

Nata, G. N. (2017). Stemming Teks Sor-Singgih Bahasa Bali. *Konferensi Nasional Sistem & Informatika 2017*.

Nuryati. (2016). *Tesaurus Bahasa Sasak.* Mataram: Kantor Bahasa Nusa Tenggara Barat.

Patil. (2017). MarS: A RuleBased Stemmer for Morphologically Rich Language Marathi. *International Conference on Computer, Communications and Electronics, pp.580–584*.

Pramudita. (2018). Klasifikasi Berita Olahraga menggunakan Metode Naive Bayes dengan Enhanced Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol.5, no.3, pp.269–276*.

Putra, R. B. (2018). Non-formal Affixed Word Stemming in Indonesian Language. *International Conference on Information and Communications Technology (ICOIACT)*.

Subali, M. A. (2019). Kombinasi Metode Rule-Based Dan N-Gram Stemming Untuk Mengenali Stemmer Bahasa Bali. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) Vol. 6, No. 2, April 2019*.

Tala, F. (2002). A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia. *Universiteit van Amsterdam*.

Widjaja, M. (2015). Implementation of Porter's Modified Stemming Algorithm in an Indonesian Word Error Detection Plugin Application. *International Journal of Technology (2015) 2: 139-150*.

Wiliam, S. (2006). Tingkat Tutur dalam Bahasa Sasak dan Bahasa Jawa. *WACANA VOL. 8 NO. 1, APRIL 2006*.