STEMMER UNTUK BAHASA MADURA DENGAN MODIFIKASI METODE *ENHANCED CONFIX STRIPPING STEMMER*

Rakhmad Maulidi

Jurusan Teknik Informatika, STIKI Malang Jl. Raya Tidar 100, Kota Malang, 65146
Telp : (0341) 560823, Fax : (0341) 562525
E-mail : maulidi.edu@gmail.com, maulidi@stiki.ac.id

***Abstrak* — Bahasa Madura memiliki morfologi yang unik dan komplek baik dari sisi sosiolinguistik, morfologi dan fonologi, dari sisi fonologi bahasa Madura memiliki fonem yang unik pada vokal dan konsonan. *Stemming* merupakan teknik ekstrasi suatu kata yang memiliki imbuhan dengan tujuan untuk mendapatkan kata dasarnya. Algoritma *Enhanced Confix Stripping Stemmer (ECS)* untuk teks berbahasa Indonesia yang memiliki tingkat keakuratan yang tinggi, algoritma ini akan dimodifikasi pada *rule base*-nya disesuaikan dengan morfologi bahasa Madura, selanjutnya akan diujicoba dengan menggunakan data uji berupa teks/puisi berbahasa Madura dan akhirnya akan dievaluasi hasilnya dari tingkat akurasi, *precison, recall* dan *F-Measure*.**

***Kata Kunci* — *Stemming*, bahasa Madura, *Enhanced Confix Stripping Stemmer*.**

I. PENDAHULUAN

Indonesia merupakan negara yang tidak hanya luas wilayahnya, akan tetapi juga kaya akan budaya, seni dan bahasa. Indonesia memiliki 707 bahasa daerah [1], sedangkan menurut Badan Bahasa Kemendiknas mencatat sejumlah 617 bahasa daerah yang tersebar di seluruh Indonesia, terdapat sejumlah 139 bahasa daerah yang terancam punah, dan sejumlah 15 bahasa daerah dinyatakan punah [2]. Ancaman tersebut kemungkinan disebabkan minimnya penggunaan bahasa daerah dalam kehidupan sehari-hari, selain itu penelitian terkait bahasa daerah kurang menarik untuk diteliti, terutama jika dikaitkan dengan teknologi informasi.

Menurut Lauder dalam [3] mengatakan bahwa bahasa Madura(bM) merupakan bahasa daerah keempat terbesar dari 13 besar bahasa daerah di Indonesia, sekitar 13,7 juta jiwa jumlah penuturnya. Bahasa Madura adalah bahasa yang komplek dan unik, baik dari sisi sosiolinguistik, morfologi dan fonologi, bahasa Madura terdiri empat dialek utama yaitu (1) dialek Sumenep, (2) dialek Pamekasan, (3) dialek Bangkalan, dan (4) dialek kangean, dan dua dialek tambahan, serta memiliki delapan tingkat tutur atau *ondhâghân bhâsa* yang terdiri dari empat tingkat tutur utama yaitu (1) *enjâ'-iyâ*, (2) *engghè-enten*, (3) *engghi-enten*, dan (4) *èngghi-bhunten*, dan empat tingkat tutur turunan [3] [4].

Keunikan lain dari bahasa Madura dari unsur fonologi khususnya dari variasi fonem, yakni pada huruf vokal dan konsonan [5]. Bahasa Madura memiliki enam buah vokal, yakni /a/, /i/, /u/, /e/, /Ə/ dan /o/ dimana keenam vokal tersebut terdiri dari lima belas alofon, alofon unik terdapat pada vokal /a/, /i/, /u/ /e/ dan /o/, sedangkan pada konsonan memiliki 31 buah dengan beberapa fonem yang unik, yaitu /'/, /bha/, /dha/, /gha/, /jha/, /th/.

Penelitian tentang *stemming* bahasa Daerah khususnya bahasa Madura masih sangat minim dilakukan, yakni dilakukan oleh Sholihin dkk [6] dengan memodifikasi metode *Enhanced Confix Stripping* yang dikembangkan oleh Arifin dkk [7], tetapi penelitian ini terbatas dalam penggunaan kata yang umum digunakan dalam dialek Bangkalan, yakni sejumlah 400 kata.

Keunikan fonem dalam bahasa Madura berpengaruh dalam pelafalan/artikulasi dan penulisan kata, kesalahan dalam penulisan kata(*typo*) khususnya penggunaan fonem yang kurang tepat akan berpengaruh pelafalan dan arti kata, banyak kata dalam bahasa Madura yang memiliki kemiripan dalam penulisan[5] diantaranya pada fonem vokal /â/ dan /a/ pada kata *rajâ*(besar) dan *raja*(raja) sedangkan pada fonem konsonan /k/ dan /'/ pada kata *katok*(bersinggungan) dan *kato'*(celana dalam), konsonan /b/ dan /bha/ pada kata *bâja* (saat, waktu) dan *bhâjâ* (buaya), konsonan /d/ dan /dha/ pada kata *dâpa'*(sampai) dan *dhâpa'*(telapak). Fonem konsonan lainnya seperti pada /g/ dan /gha/, /j/ dan /jha/ dll. Keunikan dan kemungkinan kesalahan penulisan tersebut memungkinkan menimbulkan ambiguitas, *overstemming* atau *understemming* pada saat proses stemming, misalnya pada kalimat "*Bâdâ bâddhâ beddhâ'na Raja bheddhâ rajâ ,* " (Ada wadah bedaknya raja robek besar).

Dalam rencana penelitian ini akan dikembangkan sebuah algortima stemmer untuk bahasa Madura dengan menerapkan algoritma *Enhanced Confix Stripping Stemmer* dimana *rule base* pada algoritma acuan tersebut disesuaikan

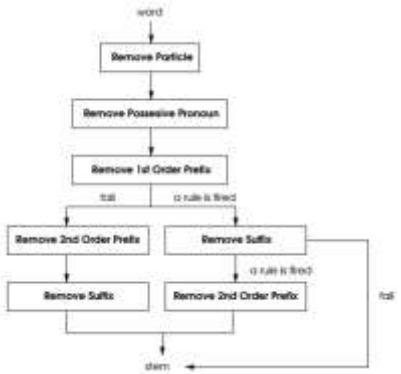
dengan morfologi bahasa Madura, dataset yang digunakan adalah kata-kata dalam dialek Sumenep, Pamekasan dan Bangkalan yakni dialek paling sering digunakan di dalam bahasa Madura, yakni kata-kata dari puisi berbahasa Madura, pemilihan puisi sebagai dataset/datauji karena variasi kata-katanya lebih beragam, terutama kata yang mengandung sisipan.

II. TINJAUAN PUSTAKA

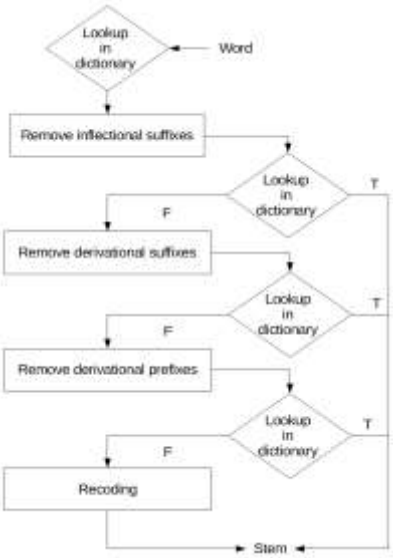
A. Stemming

Stemming merupakan proses ektraksi suatu kata dalam suatu dokumen digital yang bertujuan untuk mendapat kata dasarnya, misalnya dalam bahasa Indonesia kata 'menendang', 'tendangan', 'penendang', 'menendangi' kata dasarnya adalah 'tendang', contoh dalam bahasa Madura '*ter-penter*'(pintar-pintar), '*mamenter'*(membuat pintar), '*penterran'*(lebih pintar), '*terpenterran'*(paling pintar), '*termapenter'*(berlagak pintar) kata dasarnya adalah '*penter*"(pintar). Stemming ini nantinya bisa digunakan untuk penelitian di bidang temu kembali infomasi(*information retrieval*), misalnya translasi, ringkasan dan klasifikasi suatu dokumen[8].

Teknik *stemming* yang paling populer digunakan adalah teknik *Porter Stemmer* terutama untuk bahasa Inggris, yakni sebuah teknik stemmer dengan pendekatan *rule base* berdasarkan struktur morfologi bahasa. Teknik tersebut diadopsi oleh Tala untuk stemming bahasa Indonesia [9], teknik stemming bahasa Indonesia lainnya adalah *Config-Stripping* yang dibuat oleh Adriani dkk [8], kedua teknik stemmer bahasa Indonesia tersebut sama-sama menggunakan rule base, yang membedakan keduanya adalah penggunaan kamus bahasa Indonesia sebagai acuan, pada teknik stemmer *Config-Stripping* menggunakan kamus sebagai acuan sedangkan pada teknik Tala tidak. Perbaikan *Config-Stripping* diberi nama *Enhanced Confix Stripping Stemmer*(ECS) [7], algoritma ini memperbaiki beberapa aturan agar proses stemming berhasil untuk kata-kata dengan format "mem+p-", "mem+s-", "menge-", "penge-", "peng+k-", serta penambahan algoritma untuk mengatasi kesalahan *overstemming* yang diberi nama "iterasi pengembalian akhiran".



Gambar 1 The basic design of a Porter stemmer for Bahasa Indonesia **[9]**



Gambar 2 Design of Confix-Stripping Stemmer **[8]**

B. Morfologi Bahasa Madura

Dalam morfologi bahasa Madura banyak macamnya, menurut asal pembentukannya (*kadhâddhiânna oca'*) dibagi menjadi tiga [4], yakni:

1. Kata dasar(*oca' asal*), ialah kata belum berubah dari asalnya, contoh: *obâng*(uang), *bâto*(batu), *berraâ'*(berat), *dhâddhi*(jadi), *kakan*(makan)
2. Kata ubahan(*oca' obâ'an*), ialah kata yang sudah berubah dari asalnya setelah ditambahkan imbuhan.

Imbuhan dalam bahasa Madura diantaranya:

1. Awalan/infiks(*ter-ater*), yakni imbuhan di awal kata dasar, diantaranya: N-, a-, ta-, ka-, pa-, é-, éka, épa-, ma-, sa-, pan-, pam-, pang-, pé- contoh: *atellor*(bertelur), *épésa*(dipisah), *takébâ* (terbawa), *kaburu* (kerburu), *sabungko* (serumah), *pabhersé* (bersihkan), *panjâgâ* (penjaga), *pambâgi* (pembagi), *pangghâluy* (pengaduk), *pétotor* (penutur), *épabagus* (dibuat bagus), *ékandi'* (dimiliki), *saséba'*(sebagian).
2. Akhiran/sufiks(*panoténg*), yakni imbuhan di akhir kata dasar, diantaranya: -a, -é, -aghi, -an, -en, -na dan, contoh: *tédunga*(mau tidur), *toro'é*(ikuti), *sambiaghi* (bawakan), *nangésan* (cengeng), *kalagguen* (kepagian), *parlona* (perlunya)
3. Imbuhan dan akhiran/sufiks(*ter-ater ban panoténg*), yakni imbuhan yang diberikan pada awal dan akhir dari kata dasar, diantaranya: N-e, N-aghi, N-ana, a-e, a-aghi, ma-e, ma-an, ma-ana, ma-aghi, é-é, e-na, e-aghi, eka-e, eka-ana, eka-aghi, apa-an, epa-e, epa-aghi, ka-e, ka-aghi, pa-e, pa-aghi, pa-an, par-an, pa-na, jha-na, sa-an, sa-na, cé'-na, contoh: *ékala'aghi* (diambilkan), *katédungan* (tertidur), *akorosan* (lebih kurus), *pacolo'an* (cerewet), *ce'saké'na* (begitu sakitnya), *épaté'é* (dibunuh), *épangangguaghi*(dipakaikan)

4. Tandhuk, yakni imbuhan yang meluruhkan kata dasarnya, diantaranya: ma-, na-, nya-, nga-, contoh: *masang* (memasang), *mukka'*(membuka), *noles* (menulis), *namen* (menanam), *nyaba'* (menaruh), *nyokor* (mencukur), *ngarang* (mengarang), *ngakan* (memakan)
5. Sisipan(*sesselan*), yakni kata yang mendapatkan sisipan di tengah kata dasar misalnya: al, ar, en, om, um, contoh: *bhâ**lâ**tra*(menjadi rata), *k**ar**épé'* (terhimpit), *p**én**alang* (penghalang), *t**om**ekka*(terkabul), *gh**um**ancang*(cepat).
6. Kata ulang / Reduplikasi (*oca' rangkebbhân*) pada bahasa Madura terdiri dari tiga-macam kata ulang, yaitu: (a) Kata ulang sempurna (*Oca' rangkep buto*), yakni kata ulang dengan pengulangan penuh pada bentuk kata dasarnya, contoh: *kéra-kéra* (kira-kira), *moghâ-moghâ* (moga-moga); (b) Kata ulang depan (*Oca' rangkep ada'*), yakni kata ulang tidak sempurna yang suku kata depan diambil untuk diulang, contoh: *lalaké* (laki-laki), *papareng*(pemberian). (c) Kata ulang belakang(Oca' rangkep budhi), yakni kata ulang tidak sempurna yang suku kata belakang diambil untuk diulang, contoh: ca'oca'(kata-kata), nibenni(bukan-bukan)

III. PENELITIAN TERKAIT

Penelitian tentang stemming bahasa Indonesia sudah banyak dilakukan, yakni dengan menggunakan pendekatan *rule base* tanpa menggunakan kamus sebagai acuan [9] [10]. Penelitian lain yang menggunakan kamus sebagai acuan dilakukan oleh Nazief dan Adriani dan disempurnakan oleh Jelita Asian dengan nama *Confix-Stripping* [8]. Perbaikan *Confix-Stripping* dengan nama *Enhanced Confix Stripping Stemmer* dilakukan oleh Arifin dkk [7].

Penelitian stemming bahasa Daerah lebih sedikit dilakukan jika dibandingkan stemming bahasa Indonesia, yakni penelitian stemming bahasa Jawa dilakukan oleh Amin, dkk [11] serta Madia [12] dengan menggunakan *Rule Based Approach*, penelitian stemming bahasa Sunda oleh Junaedi dkk [13] dan Purwoko [14] dengan menggunakan *dictionary approach*, sedangkan penelitian stemming bahasa Madura dilakukan oleh Sholihin dkk [6] dengan menggunakan *Enhanced Confix Stripping Stemmer.*

IV. METODOLOGI

Metodologi dalam penelitian ini, terdiri dari empat tahapan utama, yaitu:

A. Pengumpulan data

Data yang akan disiapkan terdiri dari (a) *Dataset*, yakni data yang akan digunakan untuk menguji model stemming yang akan dikembang, data ini berasal dari kumpulan puisi berbahasa Madura, pemilihan puisi sebagai dataset karena diksi yang digunakan dalam puisi lebih bervariasi. (b) Data *Stopword*, yakni kumpulan kata dalam bahasa Madura yang kurang memiliki arti atau tidak relevan, seperti kata tanya, kata petunjuk, kata sambung, dll. kumpulan kata ini nantinya akan digunakan sebelum proses stemming dilakukan. (c) Data kamus, yakni kumpulan kata diluar data *stopword*, kumpulan kata ini terdiri dari kata kerja, kata sifat, kata benda, dll.

B. Pengembangan Algoritma Stemming

Tahapan ini merupakan inti dari penelitian ini, yakni tahap pengembangan algoritma stemming untuk bahasa Madura. Terdapat dua jenis algoritma stemmer yang dijadikan rujukan utama, yakni algoritma stemmer Tala yang tidak berbasis kamus dan algortima stemmer *Enhanced Confix Stripping Stemmer* yang berbasis kamus.

C. Pengujian

Pada tahapan ini dilakukan pengujian terhadap algoritma yang dikembangkan pada tahap sebelumnya, pengujian dilakukan dengan menggunakan sebagian dataset yang sudah disiapkan pada tahap pertama. Penggunaan sebagian data untuk pengujian bertujuan untuk mempermudah proses debuging dari algoritma yang dikembangkan, yakni menemukan kesalahan pada algoritma sehingga bisa diperbaiki sehingga menghasilkan algoritma yang paling optimal.

D. Evaluasi

Tahapan ini merupakan pengujian lanjutan dari algoritma yang dikembangkan dengan menggunakan dataset yang jauh lebih banyak dibandingkan pada tahap pengujian, pada tahap ini akan dihitung tingkat akurasinya yakni prosentase dari total kata yang benar dibandingkan jumlah total kata yang diuji coba, persamaannya sebagai berikut:

$$Akurasi = \frac{Jumlah\ Kata\ Valid\ stem}{Jumlah\ kata\ uji\ coba} \times 100\%$$

Selain itu akan dihitung nilai *precison*-nya, yaitu nilai perbandingan jumlah dokumen relevan dari hasil pencarian oleh sistem terhadap jumlah total dokumen relevan maupun tidak relevan hasil pencarian/stem, persamaannya sebagai berikut:

$$Precision = \frac{|\{relevant\ doc\} \cap \{retrieved\ doc\}|}{|\{retrieved\ doc\}|}$$

Selanjutnya akan dihitung nilai *recall*-nya yaitu perbandingan jumlah dokumen relevan dari hasil pencarian/stem terhadap jumlah total dokumen relevan dalam relevan set, persamaannya sebagai berikut:

$$Recall = \frac{|\{relevant\ doc\} \cap \{retrieved\ doc\}|}{relevant\ doc}$$

F-Measure, merupakan parameter keberhasilan retrieval yang menggabungkan nilai recall dan precision, range nilai dari F-measure antara 0 sampai 1, persamaannya sebagai berikut:

$$F - measure = 2\ x\ \frac{Precision \times Recall}{Precision + Recall}$$

Evaluasi terhadap hasil stemming berdasarkan tingkat akurasi, *precision*, *recall* dan *f-measure* merupakan pendekatan yang diajukan oleh Powers [15].

V. DISKUSI

Metode *Enhanced Confix Stripping Stemmer* mungkin bisa diterapkan untuk melakukan proses stemming teks berbahasa Madura dengan melakukan modifikasi *rule base* yang disesuaikan dengan morfologi bahasa Madura, pemilihan metode ini karena bahasa Madura masih satu rumpun dengan bahasa Indonesia [5] dan juga metode ini sudah terbukti bisa digunakan untuk stemming dokumen berbahasa Indonesia [7], dokumen berbahasa Sunda [13] [14] dan dokumen berbahasa Madura [6].

Bahasa Madura secara sosiolinguistik memiliki delapan tingkat tutur, yang terdiri dari empat tingkat tutur utama dan empat tingkat tutur turunan, serta memiliki enam dialek yang terdiri empat dialek utama dan dua dialek tambahan [3] [4], penelitian stemmer bahasa Madura yang menggunakan metode ECS yang dilakukan hanya sebatas menggunakan bahasa Madura dialek Bangkalan dan menggunakan data uji yang terbatas, yakni hanya 400 kata.

Pengujian suatu algoritma *stemmer* dengan menggunakan data uji dalam jumlah besar dan bervarisi akan menunjukkan kualitas algoritma tersebut, sehingga algortima tersebut bisa dilakukan evaluasi untuk diukur tingkat akurasinya atau *error rate*-nya secara statistik dengan menggunakan metode/pendekatan tertentu, misalnya dengan menggunakan pendekatan yang diajukan oleh Powers [15] dengan mengukur nilai akurasi, *precision*, *recall* dan *f-measure*.

Minimnya jumlah penelitian *stemming* bahasa Madura dan hanya terbatas pada satu dialek saja yang melatar belakangi rencana penelitian ini, sehingga perlu diteliti lebih dalam tentang *stemming* bahasa Madura yang bisa digunakan untuk dialek lainnya, terutama dialek utamanya yang banyak digunakan oleh masyarakat pengguna bahasa Madura, sehingga diharapkan hasil dari rencana penelitian ini yakni algoritma stemming bahasa Madura bisa dimanfaatkan untuk penelitian lainnya, misalnya untuk klasifikasi, translasi/terjemahan atau ringkasan dokumen bahasa Madura secara otomatis seperti yang dilakukan pada bahasa Indoneasia [8].

VI. KESIMPULAN

Stemming merupakan sebuah teknik yang penting dalam information retrieval, dimana teknik tersebut bisa digunakan dalam proses klasifikasi, translasi atau ringkasan dokumen secara otomatis. Bahasa Madura merupakan salah satu bahasa Daerah yang memiliki tingkat kompleksitas yang beragam terutama dari sisi morfologinya. Metode *Enhanced Confix Stripping Stemmer* memungkinkan digunakan untuk stemmer bahasa Madura dengan melakukan penyesuaian pada rule base-nya sesuai dengan morfologi bahasa Madura.

DAFTAR PUSTAKA

[1] Ethnologue.com, “Languages of Indonesia An Ethnologue Country Report,” 7 11 2016. [Online]. Available: https://www.ethnologue.com/country/ID/languages.

[2] Republika.co.id, “139 Bahasa Daerah di Indonesia Terancam Punah,” 7 11 2016. [Online]. Available: http://nasional.republika.co.id/berita/nasional/umum/16/08/02/ob9t2h383-139-bahasa-daerah-di-indonesia-terancam-punah.

[3] A. Sofyan, “Perilaku Dan Makna Verba Dalam Bahasa Madura,” *Humaniora,* pp. 333 - 344, 2012.

[4] B. dan I. Y. Fiandarti, Kosa Kata Bhasa Madhura Lengkap, Surabaya: Karya Simpati Mandiri, 2009.

[5] A. Sofian, “Beberapa Keunikan Linguistik Bahasa Madura,” *Humaniora, Volume 19*, pp. 232 - 240, 2007.

[6] A. Sholihin, F. Solihin dan F. H. Rachman, “Penerapan Modifikasi Metode Enhanced Confix Stripping Stemmer Pada Teks Berbahasa Madura,” *Jurnal Sarjana Teknik Informatika Vol. 2, No. 1,* pp. 305-324, 2013.

[7] A. Z. Arifin, I. P. A. K. Mahendra dan H. T. Ciptaningtyas, “Enhanced Confix Stripping Stemmer And Ants Algorithm For Classifying News Document In Indonesian Language,” dalam *The 5th International Conference on Information & Communication Technology and Systems*, Irbid, Jordan, 2014.

[8] M. Adriani, J. Asian, B. Nazief, S. Tahaghoghi dan H. E. Williams, “Stemming Indonesian : A Confix-Stripping Approach,” *ACMTransactions onAsian Language InformationProcessing, Vol. 6,No. 4,Article 13,* pp. 13-33, 2007.

[9] F. Z. Tala, “A Study of Stemming Effects on Information Retrieval in bahasa Indonesia,” Institut for logic, Language and Computation Universiteit van Amsterdam The Netherlands, Amsterdam, 2003.

[10] V. Vega, “Information retrieval for the Indonesian language,” National Uni- versity of Singapore, Singapore, 2001.

[11] F. Amin, Purwatiningtyas, P. Utomo, S. Ramadhanu dan S. E. Cahya, ““Stemmer Bahasa Jawa Ngoko dengan Metode Affix Removal Stemmers (Rule Based Approach),” Fakultas Teknologi Informasi Universitas Stikubank (Unisbank) Semarang., Semarang, 2016.

[12] M. Madia, “Stemming Bahasa Jawa Untuk Mencari Akar Kata Dalam Bahasa Jawa Dengan Aturan Analisis Kontrasif Afiksasi Verba,” Jurusan Teknik Informatika Fakultas Sains Dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim, Malang, 2016.

[13] D. Junaedi, I. O. Herlistiono dan D. Akbar, “Stemmer for "Basa Sunda",” dalam *Seminar Nasional Ilmu Komputer Universitas Diponegoro*, Semarang, 2010.

[14] A. Purwoko, “Model Stemming Berbasis Kamus Untuk Dokumen Berbahasa Sunda,” Sekolah Pascasarjana Institut Pertanian Bogor Bogor, Bogor, 2011.

[15] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *International Journal of Machine Learning Technology Volume 2 Issue 1,* pp. 37-63, 2011.