Pengaruh *Part of Speech Tagging* Berbasis Aturan dan Distribusi Probabilitas *Maximum Entropy* untuk Bahasa Jawa Krama

Hafiz Ridha Pramudita, Ema Utami, Armadyah Amborowati
Program Magister Teknik Informatika., Pascasarjana STMIK AMIKOM Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta
E-mail: [1]hafiz.ridha.p@gmail.com, [2]emma@nrar.net, [3]armadyah.a@amikom.ac.id

Masuk: 11 Januari 2016; Direvisi: 9 Februari 2016; Diterima: 10 Februari 2016

Abstract. *Javanese language is one of the local languages in Indonesia, which is used by most of the population of Indonesia. The language has complex grammar to embrace the values of decency that is determined by the use of words containing courtesy known as Raos Alus. Every word in the Javanese belongs to a certain part of speech like what happens to other languages. Part of Speech (POS) tagging is a process to set syntactic category in a word such as nouns, verbs, or adjectives to every word in the document or text. This study examined the POS Tagging with Maximum Entropy and Rule Based for Javanese Krama—Higher Javanese--by using the Open NLP library to measure the maximum entropy. The results obtained are Maximum Entropy and Rule Based can be used for POS Tagging on Javanese Krama with the highest accuracy of 97.67%.*
Keywords: *POS Tagging, NLP, Maximum Entropy, Rule Based, Javanese Krama Language*

Abstrak. *Bahasa Jawa merupakan salah satu bahasa daerah di Indonesia yang dipakai oleh sebagian besar penduduk Indonesia. Bahasa Jawa memiliki tata bahasa yang kompleks karena menganut nilai-nilai kesopanan yang ditentukan berdasarkan penggunaan dengan kata-kata yang mengandung raos alus (rasa sopan). Setiap kata dalam Bahasa Jawa memiliki jenis kata atau part of speech tertentu seperti halnya dengan bahasa-bahasa lain. POS tagging merupakah bagian penting dari cakupan bidang ilmu Natural Languange Processing (NLP). Penelitian ini menguji POS Tagging dengan Berbasis Aturan dan distribusi probabilitas Maximum Entropy pada Bahasa Jawa Krama menggunakan library OpenNLP untuk mengukur maximum entropy. Hasil yang diperoleh adalah Maximum Entropy dan Rule Based dapat digunakan untuk POS Tagging pada Bahasa Jawa Krama dengan akurasi tertinggi 97,67%.*
Kata Kunci: *POS Tagging, NLP, Maximum Entropy, Rule Based, Bahasa Jawa Krama*

1. Pendahuluan

Bahasa Jawa merupakan salah satu bahasa daerah di Indonesia yang dipakai oleh sebagian besar penduduk Indonesia (Quin, 2011). Telaah tentang Bahasa Jawa sendiri bukan lagi merupakan hal asing karena sebagian besar telah ditulis dalam bahasa asing (Rusyidi, dkk., 1985). Orang Jawa memiliki tata krama yang kompleks tentang bagaimana mereka berkomunikasi menggunakan bahasa dengan tingkatan yang sesuai dengan kesopanannya, sering diwujudkan dalam bentuk penghormatan yang ditunjukkan dalam komunikasi (Poedjosoedarmo, 2006). Secara garis besar ada dua tingkatan tutur dalam Bahasa Jawa, yaitu tingkat tutur kasar (*ngoko*) dan halus (*krama*). Dalam *ngoko* ada (1) *ngoko* biasa dan (2) *ngoko alus*, dalam *krama* ada (1) *krama* biasa dan (2) *krama alus*. Sampai sekarang, *unggah ungguhing* basa masih menjadi acuan dalam berbahasa Jawa (Suyata, 2011). Tingkat kesopanan tersebut dipengaruhi beberapa faktor antara lain status sosial, usia, dan hubungan persaudaraan (Sukarno, 2010). Ragam *krama* adalah bentuk *unggah-ungguh* Bahasa Jawa yang berintikan leksikon *krama*, atau yang menjadi unsur inti dalam ragam *krama*, afiks yang biasa digunakan antara lain: *dipun-*, *-ipun*, dan *–aken* (Poedjosoedarmo, 1968). Meskipun *krama* dibedakan menjadi *krama lugu* dan *krama inggil* inti dari ragam tersebut adalah leksikon yang berbentuk *karma* (Khazanah, 2012).

Part of Speech (POS) tagging adalah proses menetapkan kategori sintaksis pada sebuah kata seperti kata benda, kata ganti, kata kerja, dan kata sifat untuk setiap kata dalam dokumen atau teks. *POS tagging* merupakah bagian penting dari cakupan bidang ilmu *Natural Languange Processing (NLP)* yang diterapkan dalam pengenalan suara (*speech recognition),* pencarian informasi (*information retrieval),* pengucapan teks *(text to speech),* pencarian kata ambigu (*word sense disambiguation),* pengolahan semantik (*semantic processing),* dan mesin penerjemah (*machine translation)* (Jurafsky dan Martin, 2000).

Adapun beberapa penelitian mengenai *POS Tagging* untuk berbagai bahasa lain yang sudah dilakukan antara lain *Hidden Markov Model* (HMM) untuk Bahasa Urdu (Anwar, dkk. 2007), kemudian penelitian *POS Tagging* Bahasa Bengali yang menunjukkan akurasi *Maximum Entropy* melebihi akurasi HMM hingga 88.8% (Ekbal, dkk., 2009). Sedangkan untuk penelitian dengan berbasis aturan (*rule based)* salah satunya dalam penelitian *Rule-Based Part of Speech (RPOS) Tagging* Bahasa Malaysia yang menunjukan performa bagus dan dapat ditingkatkan akurasinya dengan penambahan *dictionary* (Alfred, dkk., 2013).

Komparasi beberapa pendekatan untuk *POS Tagging* pernah dilakukan untuk berbagai macam bahasa daerah di India seperti Bahasa Hindi, Punjabi, Malayalam, Bengali, dan Telugu dengan menggunakan *Hidden Markov Model* (HMM), *Suport Vector Machine* (SVM), *Maximum Entropy* (ME), dan *Conditional Random Field* (CRF) yang menunjukkan bahwa kekayaan morfologi suatu mempengaruhi hasil akurasi dari *POS Tagging* (Kumar dan Josan, 2010), hal ini menarik penulis untuk melakukan penelitian terhadap pengaruh *Maximum Entropy* dan *Rule-Based* pada Bahasa Jawa Krama. Pemilihan kombinasi *rule based* dan statiskal pada *POS Tagging* juga pernah digunakan dalam Bahasa Turki (Altunyurt, dkk., 2007). Sedangkan Bahasa Jawa Krama dipilih karena Bahasa Jawa Krama adalah bahasa yang digunakan dalam penulisan artikel ilmiah, berita, pidato, dan acara-acara keagamaan (Rusyidi, dkk., 1985). Penulis juga melihat perkembangan Bahasa Jawa saat ini dalam dunia teknologi yang semakin pesat antara lain tersedianya mesin penerjemah untuk Bahasa Jawa dari Google Translate serta penggunaan Aksara Jawa sudah dimasukan dalam format standar komputasi Unicode pada block U+A980–U+A9DF.

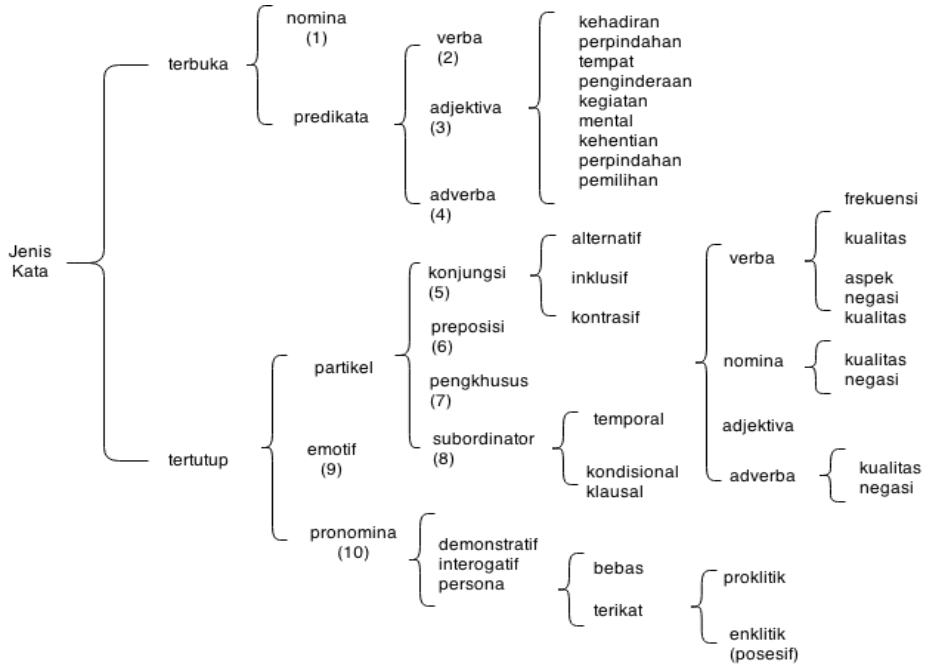
2. Landasan Teori

2.1. Jenis Kata Bahasa Jawa

Bahasa Jawa dipakai oleh sebagian besar penduduk Indonesia. Penutur asli Bahasa Jawa tidak saja menghuni sebagian besar Pulau Jawa, tetapi juga tersebar di seluruh Indonesia. Penjenisan kata dalam Bahasa Jawa dalam buku Kosa Kata Bahasa Jawa dikelompokkan menjadi tiga hal yaitu: (a) Penjenisan kata berdasar gramatikal seperti nomina dan verba. (b) Penjenisan kata berdasar kategori fungsinya seperti *ngoko, krama,* dan *krama inggil.* (c) Penjenisan kata berdasarkan sumber bahasa seperti Jawa, Indonesia, dan Inggris.

Pada Gambar 1 dapat dilihat bahwa bagan jenis kata dalam Bahasa Jawa yang telah dikelompokkan berdasarkan gramatikal. (1) Nomina adalah kata sebagai satuan leksikal yang menunjuk orang, tempat, benda, peristiwa, pengalaman, dan gagasan. Untuk membentuk suatu ungkapan arti yang lengkap berupa wujud ujaran kalimat, nomina ditaruh langsung di samping nomina lain atau di samping salah satu predikat. (2) Verba adalah jenis kata yang secara semantik menjadi pusat predikat. Kehadiran suatu verba mengharuskan adanya hubungan ketergantungan terhadapnya dari nomina yang hadir pada suatu kalimat. Verba dapat berupa keadaan maupun verba *non*-keadaan. (3) Adjektiva adalah kata yang dipakai untuk menunjuk sifat nomina. Sifat ini mengungkap ciri, keadaan, dan sifat nomina. Contoh dalam frasa *bocah ayu* (anak cantik), kata *ayu* (cantik) sebagai sifat nomina *bocah* (anak). (4) Adverba dalam Bahasa Indonesia disebut juga kata keterangan. Adverba adalah suatu kata yang secara semantik menyatakan sesuatu tentang verba, contohnya: *kanthi alon* (dengan perlahan-lahan). (5) Konjungsi atau kata sambung adalah kata untuk menghubungkan kata-kata, ungkapan-ungkapan, atau kalimat-kalimat dan sebagainya, dan tidak untuk tujuan atau maksud lain, contohnya: *utawa, apa, lan, karo.* (6) Preposisi atau kata depan adalah kata yang merangkaikan kata-kata atau bagian kalimat dan biasanya diikuti oleh nomina atau pronominal, contohnya:

saka, marang, dening, ing. (7) Pengkhusus, contohnya: *rada, banget.* (8) Subordinator disebut juga kata penghubung, yang mengubungkan induk kalimat dengan anak kalimat, contohnya: *nalika, upama, yen, jalaran, marga, supaya.* (9) Emotif adalah kata yang berhubungan dengan emosi, contohnya: *wah, adhuh, rak, lho*. (10) Pronomina atau kata ganti adalah jenis kata yang menggantikan nomina atau frasa nomina, contohnya: *iki, kui, niku, aku, kowe.*



Gambar 1. Bagan Jenis Kata Bahasa Jawa (Rusyidi, dkk., 1985**)**

2.2. *Part of Speech Tagging*

Part of Speech Tagging (POS Tagging) adalah suatu proses memberikan label kelas (anotasi) kata secara otomatis pada suatu kata dalam kalimat. Sebuah jenis kata (*part of speech)* dapat memberitahu tentang bagaimana kata tersebut dilafalkan. *POS Tagging* dapat digunakan dalam *stemming* untuk pengumpulan informasi atau yang biasa disebut dengan *Information Retrieval (IR)*, karena dengan mengetahui jenis kata dapat membantu dalam memberitahu bagian imbuhan atau afiks mana yang dapat diambil. *POS Tagging* juga dapat digunakan untuk membantu aplikasi *IR* dalam memilih kata-kata yang penting dalam dokumen, namun pada umumnya *POS Tagging* digunakan untuk '*partial parsing*' atau penguraian parsial untuk teks, misalnya untuk mempercepat pencarian nama atau frase lain untuk penggalian informasi. Algoritma untuk penandaan kata (*tagging*) tersebut secara garis besar dapat dibagi menjadi dua yaitu: penanda berbasis aturan (*rule-based taggers)* dan penanda berbasis statistik atau stochastik *(stochastic taggers)* (Jurafsky dan Martin, 2000).

2.2.1. *Rule Based Part of Speech Tagging*

Pada awal agoritma untuk otomatisasi *Part of Speech* hanya didasarkan pada dua langkah arsitektur, langkah pertama yaitu menggunakan kamus untuk menetapkan setiap kata dengan jenis kata yang potensial. Tahap kedua menggunakan daftar besar aturan disambiguasi yang ditulis secara manual untuk memisahkan setiap kata. Contoh *Rule Based POS Tagging* pada kalimat *"it isn't that odd"* dijelaskan pada Gambar 2 (Jurafsky dan Martin, 2000). Dua kalimat pertama pada aturan ini adalah untuk melihat kata "*that"* yang diikuti kalimat terakhir berupa kata sifat *(adjective)*, kata keterangan *(adverb)*, atau kata pembilang *(quantifier)* untuk kemudian dihilangkan kata keterangan *(adverb)*.

```
ADVERBIAL-THAT RULE
Given input: "that"
if
   (+1 A/ADV/QUANT); /* if next word is adj, adverb, or quantifier */
   (+2 SENT-LIM);    /* and following which is a sentence boundary, */
   (NOT -1 SVOC/A); /* and the previous word is not a verb like */
                    /* 'consider' which allows adjs as object complements */
then eliminate non-ADV tags
else eliminate ADV tag
```

Gambar 2. Contoh *Rule Based Part of Speech Tagging* (Jurafsky dan Martin, 2000)

2.2.2. *Probabilistic Part of Speech Tagging*

Probabilistics part of speech tagging merupakan salah satu pendekatan statistik (*stochastic taggers)* dalam penandaan kalimat yang pertama kali digunakan tahun 1965 yang terus mengalami perkembangan. Banyak pendekatan yang digunakan untuk mengatasi kata ambigu dan kata yang tidak diketahui atau *unknown-word*, salah satunya adalah dengan menggunakan *Maximum Entropy (MaxEnt)* atau disebut juga *multinominal logistic regression* yang pertama kali diperkenalkan oleh Ratnaparkhi (1996). *Maximum Entropy* memperkirakan probabilitas berdasarkan pada prinsip nilai asumsi terkecil, antara *data training* dan *data testing* (Ekbal, dkk., 2008).

Pemodelan statistik digunakan untuk membangun sebuah model berdasarkan sampel *output* dari proses yang merepresentasikan prediksi perilaku acak (Berger, dkk., 1996). Pada *probabilistic part of speech tagging* peluang gabungan sebuah *history h* dan anotasi (*tag*) *t* ditentukan oleh parameter-parameter yang berkaitan dengan *feature* yang aktif, yaitu α_j sehingga f_j (*h, t)*=1, α_j bertindak sebagai pembobot sebuah *feature* f_j (*h, t*). Jika diketahi (*h, t),* sebuah *feature* dapat diaktifkan pada kata atau anotasi yang berada dalam *history h*. *Feature* ini harus memiliki informasi yang dapat membantu memprediksi anotasi *t*, misalnya untuk pengejaan dari kata yang sedang dicari anotasinya atau dua anotasi sebelumnya dari kata tersebut. Sebagai contoh, dimisalkan ada sebuah *feature* dalam Bahasa Inggris seperti pada persamaan (1) (Ratnaparkhi, 1996).

Jika *feature* tersebut ada himpunan *feature* model, maka parameter model yang berkaitan akan berpengaruh terhadap peluang gabungan p(*h, t)* ketika pada kata posisi ke-*i* (w_i) berakhiran "*ing"* dan ketika anotasi kata tersebut adalah VBT (kata kerja transitif) atau t_i = VBT. Parameter α_j berpengaruh sebagai pembobot untuk memprediksi *suffix* "*ing"* untuk peluang pengamatan sebuah anotasi VBT. *Feature-feature* tersebut dibuat dengan menggunakan pasangan (*h, t)* pada data pelatihan mengikut sebuah *feature template.* Diketahui h_i Sebagai *history,* sebuah *feature* akan menanyakan berapa pertanyaan ya atau tidak mengenai h_i dan membatasi t_i pada sebuah anotasi tertentu (Ratnaparkhi, 1996).

Menurut Ratnaparkhi (1996) model peluang untuk *POS Tagging* didefinisikan atas *H*×*T*, dimana *H* adalah himpunan semua kata dan konteks anotasi yang mungkin, atau *histories,* dan *T* adalah himpunan anotasi yang diperbolehkan. Model peluang sebuah *history h* diberi anotasi *t* didefinisikan oleh persamaan (2), dimana π merupakan konstanta normalisasi, $\{\mu, \alpha_1, \ldots, \alpha_k\}$ merupakan parameter positif pembobot model, dan $\{f_1, \ldots, f_k\}$ adalah *feature,* dengan $f_j(h,t) \in \{0,1\}$. Setiap parameter α_j berkaitan dengan sebuah *feature* f_j.

Pada data pelatihan dimana terdapat urutan kata-kata $\{w_1,\ldots, w_n\}$ dan anotasi $\{t_1, \ldots, t_n\}$, h_1 didefinisikan sebagai *history* yang ada saat memprediksi t_1. Parameter pembobot $\{\mu, \alpha_1,\ldots,\alpha_k\}$ dipilih sedemikian sehingga peluang yang didapatkan dari data pelatihan (atau p) memiliki nilai setinggi mungkin. Pemilihan parameter ini dilakukan menggunakan pendekatan *conditional maximum likehood* seperti pada persamaan (3).

$$f_j(h,t) = \begin{cases} 1 & jika\ suffix(w_i) = "ing" \qquad t_i = VBT \\ 0 & selainnya \end{cases} \tag{1}$$

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_{j(h,t)}} \quad (2)$$

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_{j(h_i,t_i)}} \quad (3)$$

Model persamaan (3) dapat diinterpretasikan menggunakan formalisme *Maximum Entropy*, dengan tujuan memaksimalkan *entropy* dari distribusi tersebut dengan batasan tertentu. *Entropy* dari distribusi *p* didefinisikan seperti pada persamaan (4). *Entropy* pada persamaan (4) dibatasi oleh persamaan (5). Nilai $\tilde{E}f_i$ pada persamaan (5) didefinisikan oleh persamaan (6). Peluang (h_i, t_i) yang diamati dari data pelatihan dinyatakan dengan $\tilde{p}(t_i|h_i)$ seperti pada persamaan (6). Batasan tersebut memastikan nilai harapan *feature* dari model sesuai dengan nilai harapan yang didapatkan dari pengamatan data pelatihan (Ratnaparkhi, 1996).

$$H(p) = - \sum_{h \in H, t \in T} p(h,t) \log p\,(h,t) \quad (4)$$

$$Ef_i = \tilde{E}f_i, 1 \le j \le k \quad (5)$$

$$\tilde{E}f_j = \sum_{j=1}^{n} \tilde{p}(t_i|h_i)\, f_j(h_i, t_i) \quad (6)$$

3. Analisis Dan Perancangan Sistem

3.1. Analisis Data

Data training dan *data testing* dalam penelitian ini menggunakan data dari jurnal ilmiah BENING Volume 4, Juli 2015 yang diterbitkan oleh Universitas Negeri Yogyakarta dan telah diberikan anotasi secara manual untuk *data training* dan pengukuran akurasi. Untuk kelas kata (*tagset*) dan *dictionary* dalam penelitian ini menggunakan acuan dari buku Kosa Kata Bahasa Jawa (Rusyidi, dkk., 1985) dengan jumlah *dictionary* 10.0000 kata, dan untuk data *tagset* dilakukan pengelompokan seperti pada Tabel 1.

Tabel 1. *Data Tagset* Bahasa Jawa Modifikasi

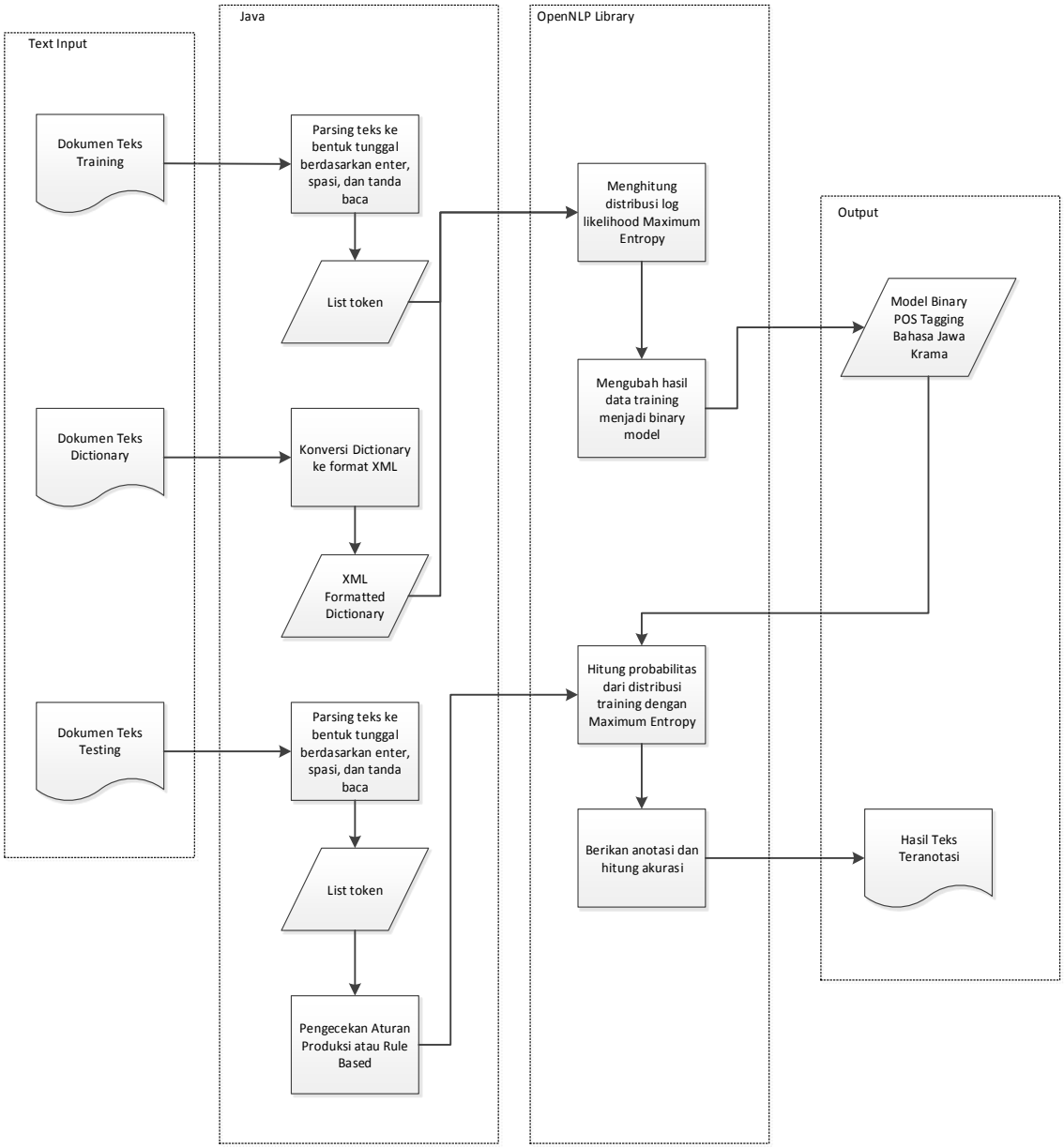
No	*Tag* Bahasa Jawa	*Penn Tree Bank Tags*	Keterangan	Contoh
1	N	NN	Nomina	toya, arta
2	V	VB	Verba	tindak, tumbas
3	Adj	JJ	Adjektiva	ayu, bagus
4	Adv	RB	Adverba	mangkih
5	Knj	CC	Konjungsi	n
6	Prp	IN	Preposisi	marang
7	Kh	-	Pengkhusus	banget
8	So	IN	Subordinator	nalika
9	Em	-	Emotif	Eh, Aduh
10	Pr	NNP / NNPS / PRP / PRP$	Pronomina	niki, niku
11	Sym	SYM	Symbol	Rp $ % & @<=> *+
12	({ [	-	Opening parenthesis	({ [
13	) }]	-	Closing parenthesis	) }]
14	,	-	Comma	,
15	‘ “	-	Quottion	‘ “
16	., ?, !	-	Sentence terminator	., ?, !
17	--, -	-	Dash	--, -
18	:	-	Colon	:
19	;	-	Semicolon	;

Modifikasi diberikan pada pelabelan tertentu seperti pada pelabelan $ dan Rp yang dimasukkan dalam *symbol* dan pelabelan tanda baca seperti yang ditunjukkan pada Tabel 1 nomor 12 sampai dengan 19. Data yang kedua adalah data untuk percobaan dan pengujian. Data percobaan diambil dari sebagian *data corpus* yang tidak memiliki pelabelan kelas katanya

sehingga hanya teks biasa. Contoh teks untuk pengujian adalah sebagai berikut: (1) *Data training*: Konflik_V sosial_N wonten_Prp ing_Prp Novel_N Piwelinge_N Putranti_N anggitanipun_V Tiwiek_N S.A_N. (2) *Data testing*: Konflik sosial wonten ing Novel Piwelinge Putranti.

3.2. Analisis Sistem

POS Tagging Bahasa Jawa Krama ini memiliki dua tahapan utama, yaitu tahapan *training* atau pelatihan dan tahapan *testing* atau pemberian anotasi. Pelatihan digunakan untuk mendapatkan model yang akan digunakan pada proses pemberian anotasi dengan menggunakan *Maximum Entropy* dan *Rule Based* pada yang dapat dilihat pada Gambar 3.



Gambar 3. Bagan Konsep Sistem *POS Tagging* Bahasa Jawa *Krama*

Proses *rule based* dilakukan ketika *testing* dimulai dengan memberikan masukan terhadap sistem. Teks masukan akan dipecah kedalam suatu kalimat dengan parameter tanda baca, kemudian dilakukan pengecekan terhadap imbuhan tertentu seperti *–ipun*, *dipun-*, *nipun-*,

dan –*aken* , setiap kata akan dilakukan pengecekan pada *dictionary* yang merupakan bagian dari *Rule Based* seperti yang ditunjukkan pada *pseudocode* Gambar 4, kemudian dilakukan perhitungan nilai probabilitas dengan menggunakan OpenNLP sebagai alat bantu untuk menghitung distribusi probabilitas *Maximum Entropy*.

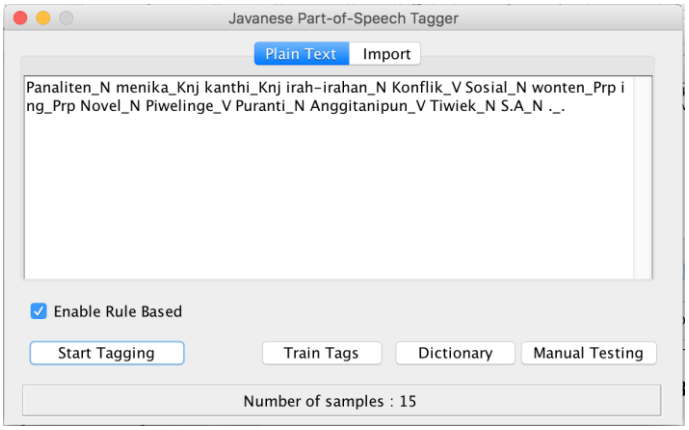
```
input teks;
String regex = "^dipun|ipun$|aken$|nipun$";
array1 = tokenisasi teks;
while semua elemen array1 mengandung regex
        if elemen ada di dictionary
                return elemen;
        else
                elemen2 = hapus awalan atau akhiran;
                if elemen2 ada di dictionary
                        return elemen2;
                else
                 return elemen;
                endif;
        endif;
endwhile;
return elemen;
```

Gambar 4. *Pseudocode* Proses *Rule-Based*

4. Implementasi dan Pengujian Sistem

4.1. Implementasi Sistem

Pada Gambar 5 diperlihatkan antarmuka *POS Tagging* Bahasa Jawa Krama. Untuk pemberian anotasi dilakukan dengan dua macam, pertama dengan masukan *plain text*, yang kedua dengan *import file*. Terdapat *check box* untuk menentukan apakah *rule based* akan diaktifkan atau tidak. Untuk melakukan proses pemberian anotasi masukkan teks yang akan diberi anotasi ke dalam *form* kemudian tekan tombol *Start Tagging*.



Gambar 5. Antarmuka *POS Tagging* Bahasa Jawa *Krama*

4.2. Hasil Pengujian

Pengujian akurasi yang dilakukan pada sistem *POS Tagging* Bahasa Jawa Krama dengan *Maximum Entropy* dan *Rule Based* dilakukan untuk *data testing* dan *data training* seperti pada Tabel 2. Masing-masing skenario pengujian dilakukan dua pengujian yaitu dengan *data training* pertama sejumlah 2380 kata (*dataset* A) dan 8488 (*dataset* B) dan 10.000 kata *dictionary*.

Tabel 2. Skenario Pengujian

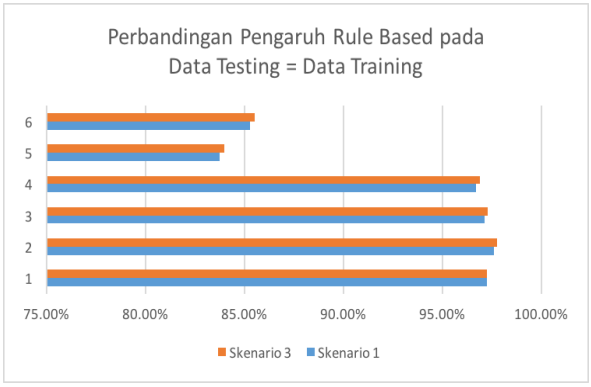
No	Data	Jenis Pengujian
1	data train = data testing	Maximum Entropy & Rule Based
2	data train ≠ data testing	Maximum Entropy & Rule Based
3	data train = data testing	Maximum Entropy
4	data train ≠ data testing	Maximum Entropy

Berdasarkan hasil pengujian dapat dilihat seperti pada grafik perbandingan akurasi antar skenario pada Gambar 6, dimana grafik tersebut menunjukkan akurasi tertinggi diperoleh pada skenario 1 dan 3, baik untuk *dataset* A maupun *dataset* B. Skenario 1B mendapatkan hasil 97,67% untuk pengujian *Maximum Entropy* dan *Rule Based* dan skenario 3B dengan akurasi tertinggi 97,85% dengan pengujian *data testing* yang merupakan bagian dari *data training* dengan pengujian *Maximum Entropy* saja. Pengujian dengan menggunakan *dataset* A menunjukkan nilai tertinggi pada skenario 2A dengan akurasi 95,75%.

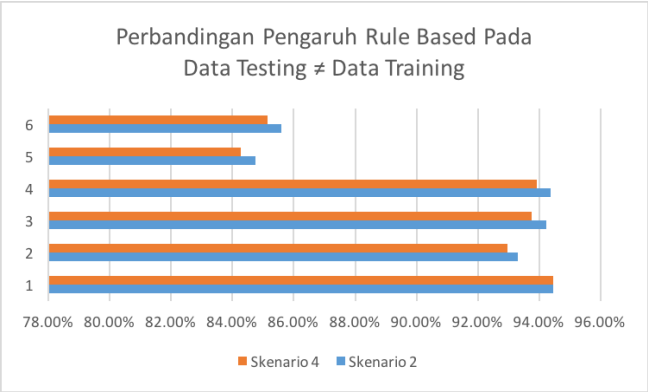
Kemudian untuk melihat pengaruh penerapan *Rule Based* dapat diperoleh dari menghitung akurasi rata-rata *dataset* A dan *dataset* B untuk tiap jenis data skenario diperoleh dari skenario 1 dan 3 yang disajikan pada Gambar 7. Skenario 1 merupakan pengujian yang dilakukan menggunakan *Maximum Entropy* dan *Rule Based*, sedangkan skenario 3 merupakan pengujian yang hanya menggunakan *Maximum Entropy*. Dari hasil yang disajikan seperti pada Gambar 7 terlihat skenario 3 memiliki akurasi yang lebih tinggi dibanding skenario 1 dengan akurasi tertinggi 97,76%. Hal ini menunjukkan perlakuan yang hanya menggunakan *Maximum Entropy* memiliki akurasi yang lebih tinggi dibanding *Maximum Entropy* dan *Rule Based* untuk *data testing* yang merupakan bagian dari *data training*.

Skenario 2 merupakan uji coba yang dilakukan dengan *Maximum Entropy* dan *Rule Based*, sedangkan skenario 4 merupakan uji coba yang hanya menggunakan *Maximum Entropy*. Berdasarkan hasil yang disajikan seperti pada Gambar 8 dapat dilihat bahwa skenario 2 memiliki hasil lebih tinggi dibandingkan dengan skenario 4 dengan akurasi tertinggi 94,37%. Hal ini menunjukkan bahwa *Maximum Entropy* dan *Rule Based* memiliki akurasi yang lebih tinggi dibandingkan yang hanya menggunakan *Maximum Entropy* saja.

Gambar 6. Grafik Tren Perbandingan Akurasi Antar Skenario



Gambar 7. Pengaruh *Rule Based* pada *Data Testing = Data Training*



Gambar 8. Pengaruh *Rule Based* pada *Data Testing ≠ Data Training*

5. Kesimpulan

Berdasarkan penelitian yang dilakukan dapat ditarik kesimpulan secara garis besar *Maximum Entropy* dan *Rule Based* dapat digunakan untuk *POS Tagging* pada Bahasa Jawa Krama dengan akurasi tertinggi 97,67% diperoleh setelah dilakukan penambahan *data training* yaitu pada pengujian skenario 1B untuk *data testing* yang merupakan bagian dari *data training* dan akurasi 95,75% untuk data yang berbeda dari *data training* pada skenario 2A. Penggunaan *Rule Based* dapat membantu ketika data *testing* merupakan data yang belum diketahui pada *data training* karena mampu meningkatkan akurasi dibandingkan hanya *Maximum Entropy* saja hingga 94,37%.

Referensi

Alfred, R., Mujat, A., & Obit, J. H. (2013). A ruled-based part of speech (rpos) tagger for malay text articles. In *Intelligent Information and Database Systems* (pp. 50-59). Springer Berlin Heidelberg.

Altunyurt, L., Orhan, Z., & Gungor, T. (2007). Towards combining rule-based and statistical part of speech tagging in agglutinative languages. *Computer engineering*, *1*(1), pp. 66-69.

Anwar, W., Wang, X., Li, L., & Wand, X. (2007). Hidden markov model based part of speech tagger for urdu. *Information Technology Journal*, pp. 1190-1198.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, *22*(1), pp.39-71.

Ekbal, A., Haque, R., & Bandyopadhyay, S. (2008). Maximum Entropy Based Bengali Part of Speech Tagging. *A. Gelbukh (Ed.), Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*, *33*, pp.67-78.

Jurafsky, D,; Martin, J, H. (2000). *Speech and language processing*, Prentice Hall, New Jersey.

Khazanah, D. (2012). Kedudukan Bahasa Jawa Ragam Krama pada Kalangan Generasi Muda, Jember, *Jurnal Pengembangan Pendidikan, 9(2).*

Kumar, D., & Josan, G. S. (2010). Part of speech taggers for morphologically rich indian languages: a survey. *International Journal of Computer Applications*, *6*(5), pp. 32-41.

Quinn, G. (2011). Teaching Javanese Respect Usage to Foreign Learners.*Electronic Journal of Foreign Language Teaching*, *8*, pp. 362-370.

Poedjosoedarmo, S. (1968). Javanese speech levels. *Indonesia* (6), pp. 54-81.

Poedjosoedarmo, G. (2006). The effect of Bahasa Indonesia as a lingua franca on the Javanese system of speech levels and their functions. *International journal of the sociology of language*, *2006*(177), pp. 111-121.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging, *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol. 1996*, pp. 133–142.

Rusyidi, Mulyanto, R.J., Sutadi, W., Suranto, Supardiman, B. (1985). *Kosa Kata Bahasa Jawa.*

Pusat Pembinaan dan Pengembangan Bahasa, Jakarta Timur.

Sukarno, S. (2010). The reflection of the Javanese cultural concepts in the politeness of Javanese. *k@ta*, *12*(1), pp. 59-71.

Suyata, P. (2011). Status Isolek Yogyakarta-Surakarta dan Implikasinya Terhadap Bahasa Jawa Standar. Litera 6(1), pp. 1-20.