SNTI2016

by Viny M

Submission date: 08-Nov-2020 07:13PM (UTC+0700)
Submission ID: 1439485927
File name: Viny-paper-snti-2016.docx (168.73K)
Word count: 3808
Character count: 23822

PART-OF-SPEECH TAGGING UNTUK BAHASA INDONESIA MENGGUNAKAN STANFORD POS-TAGGING

[1] Viny Christanti M. , M.Kom, [2] Ir. Jeanny Pragantha, M.Eng dan [3] Victor

[1,2,3] Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta
email : viny@untar.ac.id

ABSTRACT

Part-of-Speech Tagging (POS tagging) is the process of marking words in a text. Giving the class of words will depend on the relationship of the word with the next word. The relationship will be seen in the form of words, phrases, or paragraphs. The purpose of designing this application program is to implement the Stanford POS-tagging for Indonesian language.

This application is tested using 50 training document which has 14 484 words and 10 testing documents consisting of 1,897 words. Testing is done in 3 scheme. Each scheme provides different features. The highest accuracy results obtained in scheme 1 which shows the accuracy of 88.98%. In the first scheme, the number of words that obtained the correct word class is 1,688 words and the number of words that obtained wrong word class is 209 words. These results indicate that the use of appropriate features affect the outcome of the POS-tagging.

Key words

Stanford POS-Tagging, Maximum Entropy, POS Tagging

1. Pendahuluan

Dalam perkembangan teknologi, kebutuhan akan informasi menjadi semakin penting dan menjadi kebutuhan mendasar setiap manusia. Informasi dapat disampaikan melalui berbagai media salah satunya adalah melalui media elektronik. Informasi dapat disampaikan dalam berbagai bentuk seperti gambar, video atau tulisan. Setiap informasi teks disampaikan dalam berbagai bahasa. Informasi tersebut dapat dipahami oleh manusia tergantung pada bahasa masing-masing.

Bahasa adalah suatu sistem dari lambang bunyi arbitrer yang dihasilkan oleh alat ucap manusia dan dipakai untuk berkomunikasi, bekerja dan mengidentifikasikan dirinya. Bahasa yang biasa digunakan terdiri dari dua jenis yaitu bahasa lisan yang merupakan bahasa primer dan bahasa tulisan adalah bahasa sekunder [1]. Informasi tersebut dapat disampaikan dalam bentuk berita, pengetahuan umum, blog atau social media lainnya.

Penggunaan komputer dalam menyampaikan informasi sudah umum digunakan. Dengan kemajuan teknologi saat ini, memungkinkan komputer dapat memahami bahasa manusia dalam menyampaikan informasi. Penelitian dalam pengolahan bahasa untuk komputer menjadi salah satu dasar tercapainya kemampuan komputer dalam memahami bahasa manusia.

Dalam pengolahan bahasa, salah satu konsep yang perlu dipahami untuk mengolah dan menggunakan tata bahasa atau aturan yang sesuai atau yang baik dan benar adalah menentukan kelas kata. Kelas kata adalah penggolongan kata menurut bentuk, fungsi, dan maknanya. Pemberian kelas kata atau yang disebut *Part-of-speech tagging* adalah proses memberikan kelas kata kepada setiap kata dalam suatu kalimat berdasarkan konteks dari kata-kata yang berdekatan [2].

Pemberian kelas kata secara otomatis dapat dilakukan dengan berbagai macam metode seperti *Maximum Entropy*, *Rule Based* dan lain sebagainya. Berbagai penelitian sudah dilakukan untuk membuat sistem pemberian kelas kata secara otomatis. Salah satunya adalah Stanford University telah menghasilkan *POS-tagger* dengan metode *Maximum Entropy* untuk bahasa Inggris. *Stanford POS-tagger* sudah banyak diimplementasikan untuk bahasa lain seperti Arab dan Chinese [3]. *Stanford POS-tagger* merupakan salah satu aplikasi pemberi kelas kata yang akurat. Oleh karena itu pada tulisan ini dijabarkan bagaimana implementasi *Stanford POS-tagger* untuk bahasa Indonesia.

Permasalahan yang muncul dalam perancangan ini adalah bagaimana sistem *Stanford POS Tagger* melakukan training pada dokumen, menentukan jenis *tagset* bahasa Indonesia dan bagaimana sistem *Stanford POS Tagger* dapat memberikan kelas kata untuk dokumen bahasa Indonesia yang benar. Tujuan perancangan ini adalah mengimplementasi *Stanford POS tagger* untuk bahasa Indonesia dan mengetahui akurasi pemberian *POS tag* untuk bahasa Indonesia dengan metode *Maximum Entropy* berdasarkan fitur-fitur yang ada.

2. Dasar Teori

2.1 Kelas Kata dan POS-tagging

Dalam mempelajari bahasa alami terdapat beberapa tahapan yaitu fonetik atau fonologi, morfologi, sintaksis, leksikal, semantik dan pragmatik [4]. Tahapan sintaksis adalah tahapan yang berhubungan dengan pemahaman

mengenai urutan kata dalam pembentukan kalimat dan hubungan antar kata tersebut dalam proses perubahan bentuk dari kalimat menjadi bentuk yang sistematis.

Proses pemahaman kata secara sintaktis adalah mengatur tata letak suatu kata dalam kalimat yang dapat dikenali bagian-bagian kata dalam suatu kaimat yang lebih besar. Contoh pada sebuah kalimat S dapat dibentuk dari *noun phrase* (NP) dan *verb phrase* (VP). Sedangkan NP dapat berupa *determinant* (DET) atau *noun* (N). VP dapat berupa *verb* (V) atau *noun phrase* (NP). NP pada VP dapat berupa *noun* (N).

→ NP, VP (tony, makan);
NP → DET, N;
VP → V, NP;
NP → N

Dalam membentuk sebuah kalimat yang dilakukan pada tahapan sintaktis, diperlukan pengetahuan tentang kelas kata dari setiap kata. Mana kata yang termasuk sebagai kata benda atau kata kerja dan kelas kata lainnya. Proses pengenalan kelas kata dapat dilakukan menggunakan kamus bahasa.

Kelas kata atau sering juga disebut dengan jenis kata adalah pengelompokan atau penggolongan kata untuk menemukan suatu sistem dalam bahasa [5]. Kata merupakan bentuk yang sangat komplek yang tersusun atas beberapa unsur. Kelas kata terbagi menjadi lima yang didasarkan pada kategori sintaksis, fungsi, dan arti yaitu Nomina, Verba, Adjektiva, Adverbia dan Kata Tugas.

Part-of-Speech Tagging (*POS* tagging atau POST) adalah proses menandai kata-kata dalam sebuah teks yang berhubungan dengan *part-of-speech* tertentu, contohnya hubungan dengan kata-kata yang berdekatan dan terkait dalam kalimat, frase, atau paragraf [6]. *Part-of-speech* atau kelas kata adalah salah satu kelompok klasifikasi kata-kata yang sesuai fungsinya dalam suatu konteks, termasuk kata-kata seperti kata benda, kata kerja, kata sifat, kata keterangan, dan lain-lain.

POS tagging dilakukan secara manual, yaitu dengan bantuan satu atau beberapa ahli bahasa untuk memberikan *tag* yang sesuai untuk tiap kata. Namun *POS tagging* secara manual memakan banyak waktu dan biaya. Hal ini menimbulkan kebutuhan akan suatu aplikasi yang dapat memberikan atau melakukan *POS tagging* secara otomatis [8]. Metode untuk membangun sistem ini ada tiga yaitu [2]:

1. Metode *Rule Based*
   Sistem memiliki aturan pelabelan yang pengetahuannya berasal dari para ahli bahasa. Metode ini menggunakan biaya dan waktu yang sangat besar, karena pertama harus dilakukan secara manual, kedua meminta bantuan para ahli untuk membantu membuatkan *tag*, ketiga biaya yang dikeluarkan untuk membayar para ahli tersebut dan biaya operasional.
2. Metode Statistik
   Metode ini menggunakan probabilitas dari kata-kata yang muncul dalam data *training* yang kemudian akan digunakan untuk menentukan *tag* yang tepat pada kalimat yang baru. Dengan metode ini biaya dan waktu yang digunakan dapat diminimalkan, karena semuanya dapat dilakukan dalam jangka waktu yang tidak terlalu lama, serta biaya operasional yang minimal.
3. Metode *Transformation Based*
   Metode ini adalah salah satu metode berbasis korpus yang mencakup kekuatan dua dunia dapat dikatakan gabungan kedua metode sebelumnya. Metode ini mengatasi ketidakjelasan dan komplesitas tanpa mengurangi keakurasian.

2.2 *Stanford POS Tagger*

Stanford POS Tagger adalah aplikasi untuk membaca teks dan menentukan kelas kata bahasa tiap kata (dan token lainnya), seperti kata benda, kata kerja, kata sifat, dan lain-lain. Walaupun secara umum aplikasi komputasi menggunakan *POS tags* yang lebih baik seperti kata benda jamak. *Tagger* ini awalnya ditulis oleh Kristina Toutanova. Sejak saat itu Dan Klein, Christopher Manning, William Morgan, Anna Rafferty, dan Michel Galey telah meningkatkan kecepatan, kinerja, kegunaan, dan dukungan untuk bahasa lain [2].

Stanford POS Tagger menggunakan atau mengadopsi *maximum entropy approach* karena memungkinkan masuknya berbagai sumber informasi tanpa menyebabkan fragmentasi dan tanpa harus mengasumsikan kebebasan antara predictor [2]. *Tagset* yang digunakan adalah *Penn Treebank Tagset*. *Maximum Entropy* adalah rata-rata nilai informasi yang maksimum untuk suatu himpunan kejadian X dengan distribusi nilai probabilitas yang seragam [9].

Maximum Entropy Models akan menentukan kemungkinan untuk setiap tag t dalam serangkaian tag T dari kemungkinan tag kata yang diberikan dan konteksnya adalah h, yang biasanya didefinisikan sebagai urutan dari beberapa kata dan tag kata sebelumnya. Tagging adalah proses penempatan serangkaian kemungkinan maksimum tag dari serangkaian kata-kata. Model ini dapat digunakan untuk memperkirakan probabilitas dari urutan tag $t_1 \dots t_n$ diberikan kalimat $w_1 \dots w_n$ [9]:

$$P(t_1 \dots t_n | w_1 \dots w_n) = \prod_{i=1}^{n} P(t_i | t_1 \dots t_{i-1}, w_1 \dots w_n) \approx \prod_{i=1}^{n} P(t_i | h_i) \quad \dots(1)$$

dimana
$P(t_i|h_i)$ = probabilitas konteks h_i dari *tag* t_i
P = *probability* atau probabiitas
w = *word* atau kata
t = *tag*, jenis *tag* yang akan diberikan kepada konteks
h = konteks atau kata yang akan dicari probabilitasnya

Dasar pemikiran dari pemodelan maximum entropy adalah untuk memilih distribusi probabilitas p yang memiliki entropy tertinggi dari distribusi yang memenuhi kriteria dari beberapa kendala. Kendala-kendala tersebut membatasi model untuk berjalan sesuai dengan satu set statistik yang terkumpul dari data training. Data statistik dinyatakan sebagai nilai-nilai yang diharapkan dari fungsi yang sesuai yang didefinisikan pada konteks h dan tag t. Secara khusus kendala mengharapkan fitur untuk model sesuai dengan harapan empiris dari data training [2].

Contoh, jika menginginkan model untuk men-tag kata *make* sebagai kata kerja atau kata benda dengan frekuensi yang sama dengan model empiris yang disebabkan oleh data training, maka fitur didefinisikan sebagai berikut [8]:

$f_1(h,t) = 1$ *iff* $w_i = make$ *and* $t = NN$
$f_2(h,t) = 1$ *iff* $w_i = make$ *and* $t = VB$

Keterangan:
$f(h,t)$ = fitur dari konteks h *tag* t
f = *feature* atau fitur
w = *word* atau kata
t = *tag*, jenis *tag* yang akan diberikan kepada konteks
h = konteks atau kata yang akan dicari probabilitasnya

Beberapa hal umum yang menggunakan statistik untuk *part of speech tagging* adalah seberapa sering suatu jenis kata di-*tag* dalam berbagai cara, seberapa sering dua *tag* muncul berurutan atau tiga *tag* muncul berurutan. Ini terlihat seperti statistik yang digunakan oleh model Markov. Tetapi dalam kerangka maximum entropy mungkin dengan mudah dapat menentukan dan menggabungkan statistik yang lebih kompleks, tidak terbatas oleh n-gram sequence (Model n-gram sequence adalah jenis model probabilistik untuk memprediksi kata yang dicari pada urutan berikutnya [10]).

Pada model ini terdapat fitur template khusus untuk kata-kata yang jarang muncul dalam data training, untuk meningkatkan prediksi terhadap unknown words. Fitur ini adalah subset fitur yang terdapat dalam penelitian Kristina Toutanova pada tahun 2003 [2].

3. Hasil Pengujian

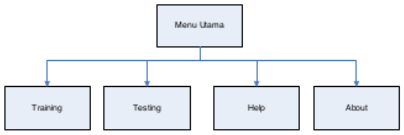
Tujuan dari perancangan ini adalah untuk melakukan *tagging* pada suatu kalimat sehingga dapat diketahui *tag* tiap kata. Data yang digunakan dalam aplikasi ini berupa kata yang sesuai dengan ejaan yang disempurnakan dalam bahasa Indonesia. Data yang diproses oleh program adalah dokumen artikel berita berbahasa Indonesia yang telah diubah ke dalam format .txt. Jumlah dokumen yang digunakan pada perancangan ini adalah sebanyak 60 dokumen artikel berita. Terdiri dari 50 dokumen artikel untuk proses *training* dan 10 dokumen artikel untuk proses pengujian

Tagset yang digunakan sebagai kelas kata pada perancangan ini adalah tagset yang sudah disesuaikan untuk bahasa Indonesia. Tabel 1 adalah daftar tagset yang digunakan pada perancangan ini.

Tabel 1 Daftar tagset yang disesuaikan untuk bahasa Indonesia Sumber: [8]

No.	Category	Postname (Tag) Berdasarkan Penn Treebank POS tagset untuk Bahasa Indonesia
1	Kata Benda/Noun (NN)	Kata Benda Tunggal (NN) Kata Benda Jamak (NNS)
2	Kata Kerja/Verb (VB)	Kata Kerja (VB)
3	Kata Sifat/Adjective (JJ)	Kata Sifat (JJ)
4	Kata Keterangan/Adverb (RB)	Kata Bantu Kerja (RB) Kata Tanya (WRB)
5	Preposition (IN)	Kata Ganti (IN)
6	Kata Penghubung/Conjunction	Kata Penghubung Setara (CC)
7	Kata Ganti/Pronoun	Kata Ganti (PRP)
8	Kata Seru/Interjection (UH)	Kata seru (UH)
9	Tanda Baca/ Punctuation (PUN)	Tanda Baca (PUN)
10	Simbol/Symbol (SYM)	Symbol (SYM)
11	Determiner (DT)	Determiner (DT)
12	Partikel/Particle (RP)	Partikel (RP)
13	Cardinal Numeral	Angka (CD)
14	Negation (NEG)	Negation (NEG)
15	Kata Asing/Foreign Word (FW)	Kata Asing (FW)

Perancangan diagram hirarki bertujuan untuk mendapatkan gambaran mengenai modul yang dibuat. Rancangan diagram hirarki dapat dilihat pada **Gambar 1**. Tampilan pertama dalam program aplikasi ini adalah menu utama yang menampilkan empat tombol menu yang mengarah ke modul-modul yang dapat dipilih pengguna, yaitu: modul *training*, modul *testing*, modul *help*, dan modul *about*.



Gambar 1 Rancangan Diagram Hirarki

Pembuatan sistem diawali dengan membuat rancangan sistem yang digunakan. Setelah itu dilakukan tahap pembuatan program aplikasi yang dimulai dari pembuatan GUI (*graphical user interface*) sampai dengan pengujian hasil dan evaluasi hasil *pos tagging* dari program yang dirancang. Program aplikasi dibuat dengan menggunakan *software* NetBeans IDE 7.0.1. Modul yang dibuat pada aplikasi ini adalah:

1. Modul Utama
   Pada Menu Utama terdapat 5 buah *button* yang dapat anda pilih untuk melakukan berbagai navigasi dalam program. *Button Testing* untuk

memanggil modul *testing*, *button Training* untuk memanggil modul *training*, *button* about untuk memanggil modul about, *button* help untuk memanggil modul help, dan *button* exit untuk mematikan aplikasi. Modul utama dapat dilihat pada Gambar 2.

Gambar 2 Modul Utama

2. Modul *Training*

Pada Modul ini proses *training* dilakukan. Modul ini berisi tentang fungsi program dalam membuat **Model** baru dengan menggunakan **fitur-fitur** yang ada dengan data training yang sudah disiapkan. Modul *training* dapat dilihat pada Gambar 3.

Gambar 3 Modul *Training*

3. Modul Testing

Pada Modul ini dapat dilakukan pengenalan *tag* atau kelas kata dengan menggunakan dok **Model** yang telah Anda buat pada proses Training**.** Modul *testing* dapat dilihat pada Gambar 4.

Gambar 4 Modul *Testing*

Tahap-tahap dalam pengujian sistem, antara lain :

1. Mengumpulkan dokumen artikel yang digunakan sebagai bahan pengujian program. Dokumen artikel didapat dari beberapa *website* seperti www.kompas.com. Artikel yang digunakan adalah sebanyak 60 buah dokumen artikel. Sebanyak 50 dokumen digunakan untuk proses *training* dan 10 dokumen digunakan untuk proses *testing*.
2. Melakukan pengujian terhadap setiap modul dan tombol untuk mengecek apakah semua modul dan tombol yang terdapat pada program berjalan dengan baik sesuai dengan fungsinya masing-masing.
3. Melakukan proses *training* untuk menghasilkan *model* yang dapat digunakan pada proses *testing* selanjutnya.

Pengujian keseluruhan terhadap aplikasi ini dilakukan dengan menjalankan Modul-Modul yang tersedia, yaitu Modul utama, Modul *training, dan* Modul *testing*, Modul hasil *testing*. Pengujian terhadap seluruh Modul dapat dikatakan berhasil karena seluruh Modul berjalan dengan baik. Semua menu dan tombol dalam masing-masing Modul dapat menjalankan fungsinya dengan baik**.**

Setelah melakukan pengujian fungsi tombol pada setiap modul, tahapan selanjutnya adalah melakukan pengujian terhadap sistem POS-tagger bahasa Indonesia secara keseluruhan. Terdapat 3 pengujian yang dilakukan. Dimana setiap fitur akan ditraining terhadap 50 dokumen yang training yang terdiri dari 14.484 kata. Pada tabel 2 dapat dilihat skema pengujian yang dilakukan. Pada setiap skema terdiri dari beberapa fitur.

Tabel 2 Daftar set Fitur yang digunakan pada pengujian

Set fitur ke-	Fitur yang digunakan
1	left3words, naacl2003unknowns
2	left3words, wordshapes(-1,1), naacl2003unknowns
3	left3words, prefix(6), suffix(6), wordshapes(-1,1), naacl2003unknowns

Contoh keterangan pada setiap fitur yang digunakan pada pengujian dapat dilihat pada tabel 3.

Tabel 3 Contoh keterangan fitur yang digunakan pada pengujian

Nama Fitur	Keterangan Fitur
left3words	Fitur untuk melakukan pemeriksaan pada kata yang dicari dan dua kata sebelumnya
prefix(6)	Fitur untuk melakukan pemeriksaan prefix atau awalan pada kalimat atau kata yang dicari
suffix(6)	Fitur untuk melakukan pemeriksaan suffix atau akhiran pada kalimat atau kata yang dicari

Setelah dilakukan training pada 50 dokumen, maka tahap berikutnya adalah pengujian dengan dokumen training. Testing dilakukan terhadap 10 dokumen artikel, dengan jumlah seluruh kata sebanyak 1.897 kata. Hasil pengujian evaluasi terhadap *tagging* dengan menggunakan tiga set fitur dapat dilihat pada **Tabel 4, 5** dan **6.**

Tabel 4 Hasil Pengujian set fitur satu

Nama	Jumlah Kata	Jumlah Kata benar	Jumlah Kata salah
dok1	149	129	20
dok2	320	292	28
dok3	156	139	17
dok4	156	145	11
dok5	209	187	22
dok6	110	103	7
dok7	290	263	27
dok8	190	163	27
dok9	206	175	31
dok10	111	92	19
Total	1.897	1.688	209

Tabel 5 Hasil Pengujian set fitur dua

Nama	Jumlah Kata	Jumlah Kata benar	Jumlah Kata salah
dok1	149	123	26
dok2	320	272	48
dok3	156	132	24
dok4	156	131	21
dok5	209	168	41
dok6	110	89	21
dok7	290	254	36
dok8	190	160	30
dok9	206	161	45
dok10	111	91	20
Total	1.897	1.585	312

Tabel 6 Hasil Pengujian set fitur tiga

Nama	Jumlah Kata	Jumlah Kata benar	Jumlah Kata salah
dok1	149	125	24
dok2	320	291	29
dok3	156	132	24
dok4	156	139	17
dok5	209	179	30
dok6	110	91	19
dok7	290	247	43
dok8	190	165	25
dok9	206	156	50
dok10	111	84	27
Total	1.897	1.609	288

Persentasi hasil pemberian *pos tagging* terhadap 10 dokumen Bahasa Indonesia dapat dilihat pada **Tabel 7**.

Tabel 7 Persentasi hasil pemberian *tag* pada dokumen bahasa Indonesia

Set fitur ke- Dokumen ke-	Fitur satu	Fitur dua	Fitur tiga
1	86,58%	82,55%	83,89%
2	91,25%	85%	90,94%
3	89,10%	84,61%	84,61%
4	92,95%	86,54%	89,10%
5	89,47%	80,38%	85,64%
6	93,64%	86,54%	90%
7	90,69%	87,59%	85,17%
8	85,79%	84,21%	87,86%
9	84,95%	78,15%	75,73%
10	82,88%	81,98%	75,67%
Total	88,98%	83,55%	84,82%

Evaluasi hasil tagging dilakukan dengan cara membandingkan secara manual hasil *tagging* dari masing-masing dokumen artikel dan hasil *tagging* dengan proses pemberian *tagging manual*. Dengan kata lain, mencocokan antara hasil *testing* dengan data *training*. Hasil dari penghitungan secara manual ini menghasilkan nilai dari tiga set fitur yang berbeda yaitu: set fitur satu sebesar 88.98%, set fitur dua sebesar 83.55%, dan set fitur tiga sebesar 84.82% untuk ketepatan pemberian *pos tagging* terhadap bahasa Indonesia dengan implementasi *Stanford POS Tagger*.

Dari hasil testing salah satu set fitur dapat diketahui penerapan aturan bahasa Indonesia pada dokumen testing dari dok10 sebagai berikut:

JAKARTA, KOMPAS.com- Pada akhir perdagangan di Nymex Sabtu dini hari tadi (17/12/2011), harga minyak mentah turun tajam. Minyak mentah Brent North Sea untuk pengiriman Februari turun 25 sen menjadi 103,35 dolar AS pada hari perdagangan pertamanya.

Harga minyak mentah merosot karena tekanan baru dari krisis utang Eropa, setelah Fitch memperingatkan Prancis kemungkinan kehilangan peringkat kredit tingkat teratasnya.

Fitch merevisi prospeknya untuk peringkat kredit Prancis menjadi negatif. Fitch Ratings pada Jumat menegaskan peringkat utang triple-A Prancis, namun merevisi prospek jangka panjang pada peringkatnya menjadi "negatif" dari "stabil.

Fitch menyatakan bahwa intensifikasi dari krisis zona euro sejak Juli merupakan sebuah kejutan negatif signifikan ke wilayah tersebut dan ekonomi Prancis serta stabilitas sektor keuangan.

Gambar 5 Dokumen dok10

Kalimat pada **Gambar 5** akan mendapatkan *tagging* secara manual seperti pada **Gambar 6**.

JAKARTA/NNP ,/, KOMPAS.com/NNP -/: Pada/RB akhir/NN perdagangan/NN di/RB Nymex/NNP Sabtu/RB dini/RB hari/NN tadi/VB -LRB-/-LRB- 17/12/2011/CD -RRB-/-RBR- ,/, harga/RB minyak/NN mentah/NN turun/NN tajam/NN ./. Minyak/NN mentah/JJ Brent/NNP North/NNP Sea/NNP untuk/TO pengiriman/NN Februari/NN turun/VB 25/CD sen/NN menjadi/RB 103,35/CD dolar/NN AS/RB pada/RB hari/RB perdagangan/NN pertamanya/NN ./. Harga/NN minyak/NN mentah/RB merosot/VB karena/RB tekanan/NN baru/JJ dari/RB krisis/NN utang/NN Eropa/NNP ,/, setelah/RB Fitch/NNP memperingatkan/VB Prancis/NNP kemungkinan/NN kehilangan/VB peringkat/NN kredit/NN tingkat/RB teratasnya/RB ./. Fitch/NN merevisi/VB prospeknya/RB untuk/TO peringkat/NN kredit/NN Prancis/NNP menjadi/RB negatif/NN ./. Fitch/NNP Ratings/NNP pada/RB Jumat/RB menegaskan/VB peringkat/NN utang/VB triple-A/NNP Prancis/NNP ,/, namun/CC merevisi/VBT prospek/NN jangka/NN panjang/JJ pada/RB peringkatnya/RB menjadi/RB ``/`` negatif/NN "/" dari/RB ``/`` stabil/NN ./. Fitch/NN menyatakan/VB bahwa/RB intensifikasi/NN dari/RB krisis/NN zona/NN euro/NN sejak/RB Juli/RB merupakan/VB sebuah/RB kejutan/NN negatif/NN signifikan/JJ ke/TO wilayah/NN tersebut/VB dan/CC ekonomi/NN Prancis/NN serta/RB stabilitas/NN sektor/NN keuangan/NN ./.

Gambar 6 Hasil *tag* secara manual

Sedangkan dari hasil pemberian *tagging* dengan program, dapat dilihat pada **Gambar 7**.

JAKARTA/NNP ,/, KOMPAS.com/NNP -/: Pada/RB akhir/NN perdagangan/NN di/RB Nymex/NNP Sabtu/NNP dini/VBI hari/RB tadi/VB -LRB-/NNP 17/12/2011/CD -RRB-/NNP ,/, harga/RB minyak/NN mentah/NN turun/NN tajam/NN ./. Minyak/VB mentah/VB Brent/NNP North/NNP Sea/NNP untuk/TO pengiriman/NN Februari/NN turun/NN 25/CD sen/NN menjadi/RB 103,35/CD dolar/NN AS/RB pada/RB hari/RB perdagangan/NN pertamanya/NN ./. Harga/CD minyak/NN mentah/RB merosot/VB karena/RB tekanan/NN baru/JJ dari/RB krisis/NN utang/NN Eropa/NNP ,/, setelah/RB Fitch/NN memperingatkan/VB Prancis/NN kemungkinan/NN kehilangan/NN peringkat/NN kredit/NN tingkat/RB teratasnya/RB ./. Fitch/NN merevisi/VB prospeknya/RB untuk/TO peringkat/NN kredit/NN Prancis/NN menjadi/RB negatif/NN ./. Fitch/NNP Ratings/NNP pada/RB Jumat/RB menegaskan/VB peringkat/NN utang/VB triple-A/NNP Prancis/NNP ,/, namun/CC merevisi/VBT prospek/NN jangka/NN panjang/JJ pada/RB peringkatnya/RB menjadi/RB ``/CD negatif/NN "/NN dari/RB ``/CD stabil/NN ./. Fitch/NN menyatakan/VB bahwa/RB intensifikasi/NN dari/RB krisis/NN zona/NN euro/NN sejak/VB Juli/NN merupakan/VB sebuah/RB kejutan/NN negatif/NN signifikan/NN ke/TO wilayah/NN tersebut/VB dan/CC ekonomi/NN Prancis/NN serta/RB stabilitas/NN sektor/NN keuangan/NN ./.

Gambar 7 Hasil *tag* program

Dari hasil testing dari salah satu set fitur dapat diketahui penerapan aturan bahasa Indonesia pada dokumen testing sebagai berikut:

Minyak mentah Brent North Sea untuk pengiriman Februari turun 25 sen menjadi 103,35 dolar AS pada hari perdagangan pertamanya.

Kalimat di atas akan mendapatkan tag secara manual sebagai berikut:

Minyak/NN mentah/RB Brent/NNP North/NNP Sea/NNP untuk/TO pengiriman/NN Februari/NN turun/VB 25/CD sen/NN menjadi/RB 103,35/CD dolar/NN AS/RB pada/RB hari/NN perdagangan/NN pertamanya/NN ./.

Sedangkan dari hasil pemberian tagging dengan program, didapat hasil:

Minyak/VB mentah/RB Brent/NNP North/NNP Sea/NNP untuk/TO pengiriman/NN Februari/NN turun/NN 25/CD sen/NN menjadi/RB 103,35/CD dolar/NN AS/RB pada/RB hari/NN perdagangan/NN pertamanya/NN ./.

Perubahan tag pada kata **turun/VB** menjadi **turun/NN** dapat disebabkan karena jumlah kata yang memiliki *tag* **NN** sebelum kata yang memiliki tag **CD** lebih banyak dibandingkan tag **VB**. Sehingga probabilitas kata mendapatkan tag **NN** lebih besar dibandingkan tag **VB**.

Kesalahan lain dapat terjadi karena struktur imbuhan yang belum sempurna untuk bahasa Indonesia. Implementasi Stanford POS-Tagging pada penelitian ini belum memperhatikan bentuk dan struktur kalimat bahasa Indonesia secara lengkap karena hanya mengadapatasi POS-tagging untuk bahasa Inggris. Namun secara keseluruhan POS-tagging dengan metode *Maximum Entropy* dari Stanford ini dapat digunakan untuk memberikan tagging kalimat bahasa Indonesia.

4. Kesimpulan dan Saran

Berdasarkan hasil pengujian yang telah dilakukan terhadap program aplikasi dan hasil ringkasan, maka dapat di tarik kesimpulan sebagai berikut:

1. Program POS Tagging Bahasa Indonesia dapat memberikan *tagging* pada dokumen bahasa Indonesia. Hal ini dibuktikan dengan hasil pengujian yang telah dilakukan terhadap 50 dokumen *training* yang memiliki 14.484 kata dan 10 dokumen testing yang terdiri dari 1.897 kata.
2. Hasil tagging terbaik diperoleh dengan menggunakan set fitur left3words dan naacl2003unknowns. Hasil *tagging* yang benar sebanyak 1.688 kata (88,98%) dan hasil *tagging* salah sebanyak 209 kata.
3. Pada saat menggunakan set fitur left3words, wordshapes(-1,1) dan naacl2003unknowns hasil *tagging* yang benar adalah sebanyak 1.585 kata (83,55%) dan hasil *tagging* salah sebanyak 312 kata.

4. Hasil *tagging* yang benar sebanyak 1.609 kata (84,82%) dan hasil *tagging* salah sebanyak 288 kata diperoleh dengan menggunakan set fitur left3words, prefix(6), suffix(6), wordshapes(-1,1), dan naacl2003unknowns.

SARAN

Saran yang dapat diberikan dalam perancangan ini adalah sebagai berikut:

1. Dalam melakukan training sebaiknya menggunakan lebih banyak jenis *tag* pada data training sehingga ketepatan dalam pemberian *tag* dapat meningkat.
2. Pengujian terhadap penggunaan fitur dilakukan kembali secara terperinci agar mendapatkan hasil yang lebih maksimal dalam pemilihan fitur untuk membangun model POS-tagger.
3. Fitur yang digunakan pada Stanford POS-tagging sebaiknya disesuaikan dengan bahasa Indonesia seperti memperbaiki daftar imbuhan yang digunakan untuk bahasa Indonesia.

REFERENSI

[1] Chaer, Abdul. *Linguistik umum*. Penerbit Rineka Cipta, 2007.

[2] Toutanova, Kristina, and Christopher D. Manning. "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger." *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics, 2000.

[3] AlGahtani, Shabib, William Black, and John McNaught. "Arabic part-of-speech tagging using transformation-based learning." *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. 2009.

[4] Smeaton, Alan F. "Natural language processing and information retrieval." *Inf. Process. Manage*. 26.1 (1990): 1-20.

[5] Kridalaksana, Harimurti. *Kelas kata dalam bahasa Indonesia*. Gramedia Pustaka Utama Cet ke-4, 2004.

[6] Pisceldo, Femphy, Ruli Manurung, and Mirna Adriani. "Probabilistic part-of-speech tagging for bahasa indonesia." *Third International MALINDO Workshop, Colocated Event ACL-IJCNLP*. 2009.

[7] Endah Purnamasari, Implementasi Program Brill Tagger Memberikan POS Tagging pada Dokomen Bahasa Indonesia, Jakarta: Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara (Skripsi tidak dipublikasikan)

[8] Chandrawati, Triastuti. "Pengembangan part of speech tagger untuk bahasa Indonesia berdasarkan metode conditional random fields dan transformation based learning." *Tersedia (online): http://www. lontar. ui. ac. id/opac/themes/libri2/d etail. jsp* (2008).

[9] The Stanford Natural Language Processing Group, *Stanford Log-linear Part-Of-Speech Tagger*, http://nlp.stanford.edu/software/tagger.shtml

[10] MacKay, David JC. Information theory, inference and learning algorithms. Cambridge university press, 2003.

Viny Christanti M., M.Kom, memperoleh gelar M.Kom dari Fakultas Ilmu Komputer, Universitas Indonesia. Saat ini aktif mengajar di Fakultas Teknologi Informasi, Universitas Tarumanagara.

Jeanny Pragantha, M.Eng., Saat ini aktif mengajar di Fakultas Teknologi Informasi, Universitas Tarumanagara.

Victor, memperoleh gelar S.Kom dari Fakultas Teknologi Informasi, Universitas Tarumanagara.

SNTI2016

ORIGINALITY REPORT

18%	15%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Brawijaya Student Paper	3%
2	research-report.umm.ac.id Internet Source	1%
3	pt.scribd.com Internet Source	1%
4	herizachaniago.blogspot.com Internet Source	1%
5	lakbanprintingmurah.com Internet Source	1%
6	repositori.kemdikbud.go.id Internet Source	1%
7	repository.unikom.ac.id Internet Source	1%
8	www.scribd.com Internet Source	1%
9	id.123dok.com Internet Source	1%

10	www.tellop.eu Internet Source	<1%
11	Submitted to Sriwijaya University Student Paper	<1%
12	text-id.123dok.com Internet Source	<1%
13	mafiadoc.com Internet Source	<1%
14	Submitted to Universitas Dian Nuswantoro Student Paper	<1%
15	Submitted to Surabaya University Student Paper	<1%
16	mozictapps.blogspot.com Internet Source	<1%
17	pdfs.semanticscholar.org Internet Source	<1%
18	id.scribd.com Internet Source	<1%
19	www.arxiv-vanity.com Internet Source	<1%
20	libraryeproceeding.telkomuniversity.ac.id Internet Source	<1%
21	Submitted to Tarumanagara University Student Paper	<1%

22	selamatdatangwelcome.blogspot.com Internet Source	<1%
23	journal.uii.ac.id Internet Source	<1%
24	"Web, Artificial Intelligence and Network Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
25	Syopiansyah Jaya Putra, Teddy Mantoro, Muhamad Nur Gunawan. "Text mining for Indonesian translation of the Quran: A systematic review", 2017 International Conference on Computing, Engineering, and Design (ICCED), 2017 Publication	<1%
26	eprints.ums.ac.id Internet Source	<1%
27	duniapendidikanversiwakamadkurikulum.blogspot.com Internet Source	<1%
28	aditawidaraputra86.blogspot.com Internet Source	<1%
29	Sifa Fauziah, Sri Muryani. "Decision Support System Untuk Menetapkan Daya Listrik Bagi Pelanggan PLN", Jurnal Perspektif, 2019 Publication	<1%

30	garuda.ristekdikti.go.id Internet Source	<1%
31	aclweb.org Internet Source	<1%
32	media.neliti.com Internet Source	<1%
33	digilib.uin-suka.ac.id Internet Source	<1%
34	ejournal.upnvj.ac.id Internet Source	<1%
35	widuri.raharja.info Internet Source	<1%
36	eprints.uny.ac.id Internet Source	<1%
37	gtarigan-gtarigan.blogspot.com Internet Source	<1%
38	bahasalami.blogspot.co.id Internet Source	<1%
39	prakerinthessa.blogspot.com Internet Source	<1%
40	www.reportshop.co.kr Internet Source	<1%
41	dblp.dagstuhl.de Internet Source	

<1%

42	"Computational Linguistics and Intelligent Text Processing", Springer Science and Business Media LLC, 2011 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		