LAPORAN TUGAS

TEXT MINING

POS TAGGER MODELING

Diajukan untuk memenuhi persyaratan kelulusan

Matakuliah IF4072 - Pemrosesan Text dan Suara Bahasa Alami

oleh :

Praditya Raudi Avinanto 13514087

Candra Ramsi 13514090

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

2017

A. Deskripsi Persoalan

Membuat Part of Speech Tagger untuk bahasa Indonesia dengan metode statistical dan menggunakan tools yang sudah ada untuk melakukan pembuatan model.

B. Dataset

Data set yang kami gunakan ialah dataset dari Universal Dependecies. Dataset tersebut berupa daftar kalimat yang kemudian diikuti dengan daftar kata beserta POS Tagnya pada kalimat tersebut. Dataset yang kami gunakan ialah id-ud-train.conllu sebagai training set dan id-ud-dev.conllu sebagai testing set. Training set kami terdiri dari 4477 kalimat dan 97531 kata unik. Sedangkan untuk testing set terdiri dari 559 kalimat dan 12612 kata unik. Sehingga total dataset yang kami gunakan terdiri dari 5036 kalimat dan 110143 kata unik. Jumlah POS Tag yang digunakan ialah 16 dengan detil sebagai berikut.

No	POS Tag	Jumlah
1	ADJ	4144
2	ADP	10866
3	ADV	4270
4	AUX	960
5	CCONJ	3284
6	DET	3594
7	NOUN	24724
8	NUM	3979
9	PART	541
10	PRON	3936
11	PROPN	20630
12	PUNCT	16504
13	SCONJ	1333
14	SYM	388
15	VERB	10952
16	X	38

Tabel 1. Detail POS Tag dan Jumlahnya Pada Dataset

Dataset tersebut menggunakan bahasa Indonesia. Kami menggunakan dataset ini karena dataset ini legal untuk digunakan.

C. Tools

Tools yang digunakan untuk proses POS Tagger modeling ini dapat dilihat pada tabel berikut

Bahasa Pemrograman	Python 2.7
IDE	Spyder
Library	Sklearn (Scikit Learn)

Tabel 2. *Tools POS Tag Modeling*

D. Tahapan

Secara garis besar proses text mining ialah menentukan feature dan training. Proses penentuan feature kami akan dijelaskan pada bagian preprocess, dan training pada bagian proses training.

a. Preprocess

Berdasarkan sumber literatur, fitur yang kami gunakan untuk pembuatan model POS Tagger terdapat 17, sebagai berikut,

1. Word : Kata yang ingin diolah
2. Is First : Kondisi dimana kata sebagai kata pertama pada kalimat
3. Is Last : Kondisi dimana kata sebagai kata terakhir pada kalimat
4. Is Capitalized : Kondisi dimana huruf pertama kata kapital
5. Is All Capitalized : Kondisi dimana semua huruf pada kata kapital
6. Is All Uncapitalized : Kondisi dimana tidak huruf pada kata yang kapital
7. Prefix-1 : Karakter pertama dari kata
8. Prefix-2 : Karakter pertama-kedua dari kata
9. Prefix-3 : Karakter pertama-ketiga dari kata
10. Suffix-1 : Karakter terakhir dari kata
11. Suffix-2 : Karakter terkahir-kedua terakhir dari kata
12. Suffix-3 : Karakter terakhir-ketiga terakhir dari kata
13. Prev Tag : Tag pada kata sebelumnya
14. Prev Word : Kata sebelumnya
15. Has Hyphen : Kondisi dimana kata memiliki karakter ‘-’
16. Is Numeric : Kondisi dimana kata berbentuk numerik

17. Is Capital Inside : Kondisi dimana terdapat huruf kapital pada salah satu karakter kecuali karakter pertama kata.

Preprocessing yang kami lakukan hanyalah mengolah data raw dari dataset menjadi data yang terdiri dari 17 fitur diatas.

b. Proses Training

Berdasarkan hasil studi literatur didapatkan tiga classifier terbaik yang diugnakan pada proses training kami, yaitu,

1. Multi Layer Perceptron
2. Stochastic Gradient Descent
3. Support Vector Classification

Semua *classifier* yang digunakan diimplementasikan menggunakan *library* Sklearn dengan masing-masing konfigurasi sebagai berikut,

Classifier	Parameter	Value
Multi Layer Perceptron	Jumlah Hidden Layers	4
	Random State	1
	Alpha	2e-1*5 penalty (regularization term) parameter
	Solver Algorithm	‘lbfgs’ is an optimizer in the family of quasi-Newton methods.
	Neutron Per Layer	17
	Epoch	200
Stochastic Gradient Descent	Loss	Perhitungan loss function menggunakan logistik regression
Support Vector Classification	Decision Function Shape	Menggunakan one vs one decision function

Tabel 3. Konfigurasi Classifier

E. Hasil dan Analisis

Pada eksperimen yang kami lakukan, kami mencoba tiga *classifier* pada lima jenis *dataset*. Lima *dataset* ini dibagi berdasarkan fitur yang digunakan. Kami mencoba membuat variasi fitur dari 17 fitur diatas. Penentuan variasi fitur ini didasarkan pada hipotesis kami dari sudut pandang pengaruh fitur tersebut dalam proses pembuatan model dan juga studi literatur. Berikut merupakan tabel variasi fitur yang kami gunakan,

ID	Fitur Yang Dihapus Dari 17 Fitur Diatas
1	Is Capital Inside, Has Hyphen
2	Prev Tag
3	Prefix, Suffix
4	Is Capitalized
5	-

Tabel 4. Kombinasi Fitur Untuk Eksperimen

Evaluasi eksperimen ini menggunakan nilai akurasi dan juga f1 score. Pada perhitungan f1 score kami menggunakan Sklearn dengan parameter weighted average dikarenakan data kami terdiri dari banyak kelas. Berikut merupakan hasil eksperimen yang telah kami lakukan,

No	Classifier	Feature Combination ID	F1 Score	Accuracy	Average Accuracy
1	SGD	1	0.9096	0.9111	0.8966
		2	0.9047	0.9067	
		3	0.8419	0.8484	
		4	0.9066	0.9083	
		5	0.9065	0.9086	
2	MLP	1	0.9140	0.9142	0.9076
		2	0.9186	0.9185	
		3	0.8758	0.8750	
		4	0.9160	0.9163	
		5	0.9142	0.9142	

3	SVM	1	0.4166	0.5232	0.4939
		2	0.4175	0.5247	
		3	0.4121	0.5312	
		4	0.2604	0.3673	
		5	0.4166	0.5232	

Tabel 5. Hasil Eksperimen

Dapat dilihat dari tabel diatas, untuk *classifier* SGD dan MLP memberikan hasil yang cukup baik yakni dengan akurasi rata-rata 89% dan 90%. Dengan demikian dapat disimpulkan bahwa MLP merupakan *classifier* dengan akurasi terbaik diantara ketiga classifier yang ada. Selain itu perlu diperhatikan bahwa pada SGD maupun MLP, hilangnya *suffix* dan prefix cukup memberikan hasil yang signifikan yakni akurasi yang turun cukup drastis. Hal ini membuktikan bahwa prefix dan suffix memberikan pengaruh cukup penting dalam POS Tag bahasa Indonesia, karena prefix dan suffix ialah lambang dari imbuhan kata yang dapat membedakan kata kerja dan kata benda. SVM memberikan hasil terburuk yakni dengan akurasi rata-rata 49%. SVM juga membutuhkan waktu relatif lama untuk training yakni 8 hingga 10 kali lebih lambat dibanding kan dengan MLP. Fitur yang paling berpengaruh pada SVM diantara lima variasi fitur tersebut ialah is capitalized, karena telah menurunkan rata-rata akurasi dari kelima dataset lain. Sedangkan untuk nilai F1 Score keseluruhan, cukup mendekati nilai akurasi.

Dari hasil tersebut kami memutuskan untuk menggunakan eksperimen dengan akurasi terbaik yakni classifier MLP dengan fitur variasi kedua untuk pembuatan model POS Tagger kedepannya. Berikut merupakan hasil percobaan model POS Tagger kami terhadap input dari user.

```
Masukkan kalimant : Raudi dan Candra mengumpulkan kode program POS tagger yang bisa menerima
masukan kalimat dan mengeluarkan keluaran  berupa daftar POS tag untuk kalimat tersebut.
List kata
['Raudi', 'dan', 'Candra', 'mengumpulkan', 'kode', 'program', 'POS', 'tagger', 'yang', 'bisa',
'menerima', 'masukan', 'kalimat', 'dan', 'mengeluarkan', 'keluaran', 'berupa', 'daftar', 'POS',
'tag', 'untuk', 'kalimat', 'tersebut', '.']

List Tag
['PROPN' 'CCONJ' 'PROPN' 'VERB' 'NOUN' 'NOUN' 'PROPN' 'NOUN' 'PRON' 'ADV' 'VERB' 'VERB' 'NOUN'
'CCONJ' 'VERB' 'NOUN' 'VERB' 'NOUN' 'PROPN' 'NOUN' 'ADP' 'NOUN' 'DET' 'PUNCT' ]
```

Gambar 1. Hasil Percobaan POS Tagger Terhadap Input User

Dapat dilihat dari hasil percobaan diatas, POS Tagger kami sudah memberikan hasil yang cukup baik.

F. Simpulan

POS Tagger modeling dapat dilakukan menggunakan classifier MLP dan SGD namun kurang baik jika menggunakan SVM. Fitur yang cukup berperan diantara lima variasi fitur ialah prefix dan suffix serta is capitalized. Model terbaik ialah classifier MLP dengan dataset variasi fitur kedua. Menggunakan model tersebut, kami dapat melakukan POS Tagging dengan baik.