

# LAPORAN TUGAS PEKAN 13 NLP : EKSTRAKSI INFORMASI

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

## 1. Deskripsi Masalah

Pengamatan eksperimen Named Entity Recognition (NER), metode yang digunakan adalah CRF, di mana digunakan beberapa fitur.

Akan dilakukan eksperimen dengan menggunakan fitur yang berbeda:

1. Tanpa fitur POSTag current kata dan POSTag kata konteks

2. Tanpa fitur ortografi (terlihat dari fungsi istitle(), isupper(), dan isdigit()) current kata dan pada kata konteks

3. Tanpa fitur akhiran pada current kata

Kemudian akan dibandingkan hasilnya dengan penggunaan keseluruhan fitur berdasarkan nilai precision, recall, dan f1-score.

## b. Tanpa fitur POSTag

	precision	recall	f1-score	support
B-art	0.00	0.00	0.00	29
B-eve	1.00	0.17	0.29	24
B-geo	0.76	0.87	0.81	1043
B-gpe	0.87	0.78	0.82	588
B-nat	0.80	0.40	0.53	10
B-org	0.74	0.61	0.67	649
B-per	0.81	0.77	0.79	546
B-tim	0.91	0.83	0.87	589
I-art	0.00	0.00	0.00	7
I-eve	1.00	0.11	0.20	18
I-geo	0.68	0.71	0.70	204
I-gpe	0.50	0.53	0.51	17
I-nat	1.00	0.50	0.67	2
I-org	0.78	0.73	0.75	545
I-per	0.79	0.89	0.83	574
I-tim	0.79	0.63	0.70	185
micro avg	0.79	0.77	0.78	5030
macro avg	0.71	0.53	0.57	5030
weighted avg	0.79	0.77	0.77	5030

## c. Tanpa fitur ortografi

	precision	recall	f1-score	support
B-art	0.33	0.03	0.06	29
B-eve	0.50	0.17	0.25	24
B-geo	0.72	0.87	0.79	1043
B-gpe	0.85	0.72	0.78	588
B-nat	0.00	0.00	0.00	10
B-org	0.74	0.59	0.65	649
B-per	0.80	0.80	0.80	546
B-tim	0.88	0.76	0.82	589
I-art	0.00	0.00	0.00	7
I-eve	0.44	0.22	0.30	18
I-geo	0.66	0.71	0.68	204
I-gpe	0.67	0.47	0.55	17
I-nat	0.00	0.00	0.00	2
I-org	0.75	0.70	0.72	545
I-per	0.79	0.90	0.84	574
I-tim	0.69	0.61	0.64	185
micro avg	0.77	0.75	0.76	5030
macro avg	0.55	0.47	0.49	5030
weighted avg	0.76	0.75	0.75	5030

## 2. Hasil dan Analisis

### a. Menggunakan semua fitur

	precision	recall	f1-score	support
B-art	1.00	0.03	0.07	29
B-eve	0.86	0.25	0.39	24
B-geo	0.75	0.88	0.81	1043
B-gpe	0.89	0.78	0.83	588
B-nat	0.67	0.20	0.31	10
B-org	0.75	0.64	0.69	649
B-per	0.81	0.81	0.81	546
B-tim	0.90	0.85	0.87	589
I-art	0.00	0.00	0.00	7
I-eve	0.57	0.22	0.32	18
I-geo	0.71	0.71	0.71	204
I-gpe	0.47	0.53	0.50	17
I-nat	1.00	0.50	0.67	2
I-org	0.78	0.73	0.76	545
I-per	0.80	0.90	0.85	574
I-tim	0.79	0.68	0.73	185
micro avg	0.80	0.78	0.79	5030
macro avg	0.73	0.54	0.58	5030
weighted avg	0.80	0.78	0.78	5030

## LAPORAN TUGAS PEKAN 13 NLP : EKSTRAKSI INFORMASI

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

### d. Tanpa fitur akhiran kata

	precision	recall	f1-score	support
B-art	1.00	0.03	0.07	29
B-eve	1.00	0.25	0.40	24
B-geo	0.74	0.88	0.81	1043
B-gpe	0.88	0.76	0.81	588
B-nat	0.67	0.20	0.31	10
B-org	0.74	0.62	0.68	649
B-per	0.82	0.80	0.81	546
B-tim	0.92	0.83	0.87	589
I-art	0.00	0.00	0.00	7
I-eve	1.00	0.22	0.36	18
I-geo	0.69	0.71	0.70	204
I-gpe	0.45	0.53	0.49	17
I-nat	1.00	0.50	0.67	2
I-org	0.78	0.73	0.76	545
I-per	0.80	0.91	0.85	574
I-tim	0.81	0.68	0.74	185
micro avg	0.79	0.78	0.78	5030
macro avg	0.77	0.54	0.58	5030
weighted avg	0.80	0.78	0.78	5030

Dari ketiga nilai yang dihasilkan, eksperimen tanpa menggunakan ortografi mendapatkan nilai terendah. Hal ini dapat terjadi karena dataset yang diolah adalah teks formal, yang dimana huruf kapital atau angka atau suatu judul mempengaruhi makna dari suatu kata atau frasa. Kemudian untuk fitur akhiran kata dan fitur POSTag tidak terlalu berpengaruh, karena dari ketiga nilai tersebut hanya terdapat sedikit perbedaan pada nilai yang menggunakan semua fitur.

Dilihat dari nilai macro average f1-score, nilai terendah dihasilkan dari eksperimen tanpa fitur ortografi dengan nilai 0,49. Nilai tertinggi dihasilkan dari eksperimen menggunakan semua fitur dan tanpa fitur akhiran kata dengan nilai 0,58.

Dilihat dari nilai macro average precision, nilai terendah dihasilkan dari eksperimen tanpa fitur ortografi dengan nilai 0,55. Nilai tertinggi dihasilkan dari eksperimen tanpa fitur akhiran kata dengan nilai 0,77.

Dilihat dari nilai macro average recall, nilai terendah dihasilkan dari eksperimen tanpa ortografi dengan nilai 0,47. Nilai tertinggi dihasilkan dari eksperimen menggunakan semua fitur dan tanpa fitur akhiran kata dengan nilai 0,54.