Register : Jurnal Ilmiah Teknologi Sistem Informasi 4 (2) 2018 81-90
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Tersedia *online* di www.journal.unipdu.ac.id
Unipdu

Halaman jurnal di www.journal.unipdu.ac.id/index.php/register

Klasifikasi jenis kejadian menggunakan kombinasi NeuroNER dan *Recurrent Convolutional Neural Network* pada data Twitter

Fatra Nonggala Putra [a], *Chastine Fatichah* [b]

[a,b] *Teknik Informatika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*

email: [a] *putra.fatra08@gmail.com*, [b] *chastine@cs.its.ac.id*

INFO ARTIKEL

Sejarah artikel:
Menerima 27 Juli 2018
Revisi 2 Agustus 2018
Diterima 8 Agustus 2018
Online 12 Agustus 2018

Kata kunci:
deteksi kejadian
ekstraksi informasi
NeuroNER
RCNN

Keywords:
incident detection
information extraction
NeuroNER
RCNN

Style APA dalam mensitasi artikel ini:
Putra, F. N., & Fatichah, C. (2018). Klasifikasi jenis kejadian menggunakan kombinasi NeuroNER dan Recurrent Convolutional Neural Network pada data Twitter. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 4(2), 81-90.

ABSTRAK

Sistem deteksi kejadian dari data Twitter bertujuan untuk mendapatkan data secara *real-time* sebagai alternatif sistem deteksi kejadian yang murah. Penelitian tentang sistem deteksi kejadian telah dilakukan sebelumnya. Salah satu modul utama dari sistem deteksi kejadian adalah modul klasifikasi jenis kejadian. Informasi dapat diklasifikasikan sebagai kejadian penting jika memiliki entitas yang merepresentasikan di mana lokasi kejadian terjadi. Beberapa penelitian sebelumnya masih memanfaatkan fitur 'buatan tangan', maupun fitur model berbasis *pipeline* seperti *n-gram* sebagai penentuan fitur kunci klasifikasi yang tidak efektif dengan performa kurang optimal. Oleh karena itu, diusulkan penggabungan metode *Neuro Named Entity Recognition* (NeuroNER) dan klasifier *Recurrent Convolutional Neural Network* (RCNN) yang diharapkan dapat melakukan deteksi kejadian secara efektif dan optimal. Pertama, sistem melakukan pengenalan entitas bernama pada data *tweet* untuk mengenali entitas lokasi yang terdapat dalam teks *tweet*, karena informasi kejadian haruslah memiliki minimal satu entitas lokasi. Kedua, jika *tweet* terdeteksi memiliki entitas lokasi maka akan dilakukan proses klasifikasi kejadian menggunakan klasifier RCNN. Berdasarkan hasil uji coba, disimpulkan bahwa sistem deteksi kejadian menggunakan penggabungan NeuroNER dan RCNN bekerja dengan sangat baik dengan nilai rata-rata *precision*, *recall*, dan *f-measure* masing-masing 94,87%, 92,73%, dan 93,73%.

ABSTRACT

The incident detection system from Twitter data aims to obtain real-time information as an alternative low-cost incident detection system. One of the main modules in the incident detection system is the classification module. Information is classified as important incident if it has an entity that represents where the incident occurred. Some previous studies still use 'handmade' features as well as feature-based pipeline models such as n-grams as the key features for classification which are deemed as ineffective. Therefore, this research propose a combination of Neuro Named Entity Recognition (NeuroNER) *and Recurrent Convolutional Neural Network* (RCNN) *as an effective classification method for incident detection. First, the system perform named entity recognition to identify the location contained in the tweet text because the event information should have at least one location entity. Then, if the location is successfully identified, the incident will be classified using* RCNN. *Experimental result shows that the incident detection system using combination of* NeuroNER *and* RCNN *works very well with the average value of precision, recall, and f-measure 92.44%, 94.76%, and 93.53% respectively.*

1. Pendahuluan

Penggunaan sosial media untuk berbagi informasi sudah menjadi kebiasaan masyarakat pengguna intenet. Salah satu media sosial yang memproduksi kiriman terbanyak adalah Twitter. Twitter dapat

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

menghasilkan rata-rata sebanyak 320 juta *tweet* perhari atau sekitar 3.700 *tweet* perdetik dengan pengguna aktif mencapai 317 juta setiap bulannya (Nidhi & Annappa, 2017). Maka tidak salah jika Twitter disebut sebagai *microblogging* terpopuler di dunia saat ini. Selain sebagai ajang curhat kaum milenial umumnya, banyak pengguna dari berbagai kalangan memanfatkan Twitter sebagai ajang berbagi informasi seputar masalah-masalah yang ada di sekitarnya. Hal ini dimanfaatkan dalam beberapa penelitian yang sudah ada sebagai data untuk deteksi komunitas tersembunyi dalam media sosial media (He, Li, Soundarajan, & Hopcroft, 2018), analisa sentimen sarkastik (Bharti, Vachha, Pradhan, Babu, & Jena, 2016), penentuan jalur optimal (Hasby & Khodra, 2013), dan deteksi kejadian lalu-lintas (Gu, Qian, & Chen, 2016).

Dari penelitian yang sudah ada, beberapa penelitian dilakukan untuk mendeteksi kejadian dari data Twitter, karena dianggap lebih murah dari pada deteksi kejadian konvensional menggunakan berbagai sensor yang dipasang pada titik lokasi diseluruh penjuru kota (Gu, Qian, & Chen, 2016). Gu, Qian, dan Chen (2016) menggunakan klasifier Naive Bayes untuk melakukan klasifikasi biner pada data *stream* Twitter sebagai *tweet* informatif, *tweet* noninformatif yang kemudian dilakukan *geoparsing* untuk mendapatkan lokasi kejadian dan diklasifikasikan lagi menjadi lima kelas jenis kejadian. Sedangkan pada Hasby dan Khodra (2013) menggunakan data *tweet* yang berupa informasi kejadian lalu-lintas untuk pencarian rute optimal guna menghindari lokasi kejadian yang telah dideteksi oleh sistem. Pada Perdana, Fatichah, & Purwitasari (2015) melakukan deteksi kejadian yang berulang secara periodik (*trivial*) dengan menggunakan metode *Autocorrelation Wavelet Coefficients*. Namun, pada penelitian tersebut tidak memperhitungkan terdapatnya entitas lokasi sebagi unsur utama suatu informasi kejadian penting.

Salah satu proses terpenting dalam sistem deteksi kejadian adalah proses klasifikasi. Proses klasifikasi digunakan untuk menentukan sebuah *tweet* sebagai kelas *tweet* noninformasi kejadian, ataukah merupakan *tweet* informasi kejadian seperti kelas kejadian lalu-lintas, kebakaran, dan gempa. Sedangkan syarat utama dari suatu informasi dapat dikatakan sebagai informasi kejadian penting jika memiliki entitas lokasi kejadian di dalamnya. Maka dari itu, selain klasifikasi proses terpenting lainnya adalah pengenalan entitas bernama pada teks *tweet* atau disebut sebagai *named-entity-recognition* (NER). Prinsip dari NER serupa dengan POS-*tagging* yaitu klasifikasi kelas kata pada teks seperti kata kerja, kata sifat, dan lain-lain (Najibullah & Mingyan, 2015).

Beberapa pendekatan dalam penentuan nama lokasi dalam sebuah *tweet* terdiri dari enam, yaitu: Penentuan lokasi berbasis klasifikasi kata dan atau frase bertipe *noun*, model bahasa, pencocokan *gazetteer*, asosiasi terhadap nama tempat yang di-*tag*, koordinat geografis pengguna, dan berdasarkan singkatan nama tempat (Gelernter & Balaji, 2013). Model dari NER sendiri terus berkembang dengan variasinya. Salah satu model terbaru dari NER adalah dengan menggabungkan metode NER konvensional dengan arsitektur *Neural Network* (Lai, Xu, Liu, & Zhao, 2015). Kemudian (Dernoncourt, Lee, & Szolovits, 2017) membangun program implementasi dari arsitektur *Neural Network* dan CRF untuk pengenalan entitas bernama dengan penamaan metode yang disebut sebagai NeuroNER.

Data Twitter yang berupa *tweet* memiliki beberapa keunikan khusus dibandingkan data teks dokumen *online* lainnya, seperti berita maupun artikel. Beberapa keunikan dari data *tweet* yang menjadi masalah pada pemrosesan data teks antara lain: Penggunaan akronim yang tidak baku, penulisan kata diganti dengan angka seperti 'perempatan' ditulis 'per4an', pengulangan huruf pada kata seperti 'macet' ditulis 'maaacceeeett', dan banyak lainnya. Guna mengatasi masalah-masalah yang ada dalam data *tweet* diperlukan jenis praproses yang bisa mengatasi keunikan yang ada dalam data *tweet* tanpa mengurangi informasi yang ditulis dalam sebuah *tweet*.

Oleh karena itu, pada penelitian diusulkan penggunaan NeuroNER dan RCNN untuk mengetahui performa keduanya jika diterapkan pada sistem deteksi kejadian dengan menggunakan data Twitter yang mempunyai banyak keunikan dibandingkan data dokumen *online* lainnya. NeuroNER sebagai pengenal entitas bernama berguna sebagai *flag* yang dapat mengurangi input ke proses klasifikasi karena mengharuskan data mempunyai paling tidak satu entitas representasi lokasi sebagai syarat utama dari informasi kejadian penting. Klasifier RCNN dengan segala kelebihannya dibandingkan model *neural* lainnya diharapkan dapat bekerja dengan baik, meski harus harus berhadapan dengan data Twitter.

ISSN 2502-3357 (*Online*)
F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90 ISSN 2503-0477 (*Print*)

2. *State of the Art*

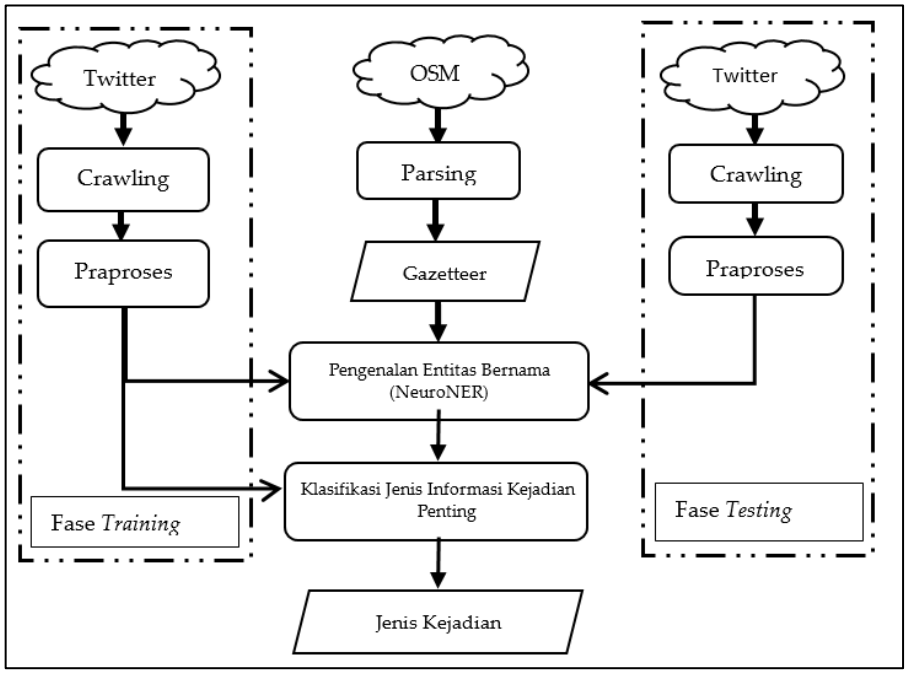
2.1. *Teks praproses*

Praproses dalam pemrosesan teks sangat berperan penting agar data siap untuk diproses yang juga akan meningkatkan hasil dari penelitian (Putra, Effendi, & Arifin, 2018). Keunikan data Twitter dapat diatasi dengan diterapkan beberapa jenis praproses yang sesuai dengan karakteristik data tersebut. Berdasarkan pada (Jianqiang & Xiaolin, 2017), mengubah singkatan ke bentuk panjangnya, penghapusan URL dan angka pada teks Twitter meningkatkan performa dari klasifikasi yang menunjukkan bahwa URL tidak berisi informasi yang berguna sebagai fitur klasifikasi. Namun, URL juga tidak menaikkan nilai *precision* dan *recall* sistem, hanya saja penghapusan URL akan mengurangi ukuran fitur data klasifikasi.

2.2. *Pengenalan entitas bernama*

Penggunaan NER konvensional pada data *tweet* telah dilakukan oleh (Liu & Zhou, 2013) yang menghasilkan hasil yang kurang memuaskan. Lample, Ballesteros, Subramanian, Kawakami, dan Dyer (2016) dan Huang, Xu, dan Yu (2015) melakukan perbaikan pada model *Named Entity Recognition* (NER) lama, yaitu dengan memperkenalkan arsitektur *neural network* berbasis *bi-Directional Long Short Term Memory* (bi-LSTM) dan *Conditional Random Fields* (CRF) untuk mengenali empat tipe entitas bernama: Lokasi, orang, organisasi, dan *miscellaneous* yang bekerja optimal pada dokumen *online* dibandingkan NER konvensional. Berikutnya, Dernoncourt, Lee, dan Szolovits (2017) memperkenalkan nama NeuroNER sebagai representasi penggunaan arsitektur *Neural Network* pada NER yang menunjukkan bahwa model NeuroNER berkerja lebih baik daripada model NER konvensional yang sudah ada. Namun, kedua model *neural* NER tersebut belum diuji pada data Twitter seperti pada (Liu & Zhou, 2013).

2.3. *Teks klasifikasi*



Gambar 1. Desain sistem deteksi kejadian

Lai, Xu, Liu, dan Zhao (2015) memperkenalkan *Recurrent Convolutional Neural Network* sebagai pengembangan dari *Recurrent Neural Network* (RNN) dan *Convolutional Neural Network* (CNN) dengan memperbaiki kekurangan dan menggunakan kelebihan dari kedua algoritma tersebut. Masih dalam Lai, Xu, Liu, dan Zhao (2015), RCNN menggunakan *bi-Directional Recurrent* struktur yang mempunyai kemampuan menyeleksi fitur yang tidak penting atau *noise* dibandingkan model *Neural Network* tradisional berbasis *window*. Karena penggunaan *max-pooling layer* akan secara otomatis menilai semua fitur secara keseluruhan dan memilih fitur terbaik yang memiliki bobot tertinggi sebagai kata kunci dalam klasifikasi teks.

F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

3. Metode Penelitian

Sistem deteksi kejadian yang peneliti bahas hanya berfokus pada tahap penentuan jenis kejadian pada klasifikasi jenis kejadian pada sistem deteksi kejadian yang telah diusulkan oleh (Gu, Qian, & Chen, 2016). Berdasarkan penelitian tersebut peneliti merumuskan model baru dalam deteksi kejadian dengan menggunakan NeuroNER dan RCNN serta menghapus proses klasifikasi biner pada sistem. Desain sistem secara keseluruhan ditunjukkan pada Gambar 1.

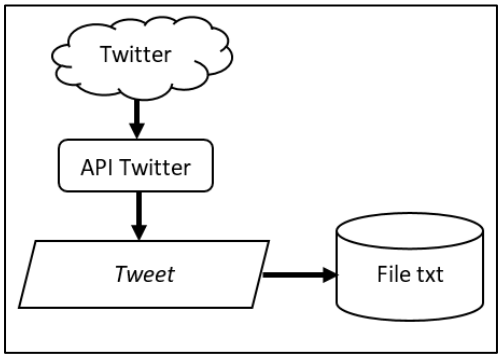
3.1. Pengumpulan data

Ada dua jenis data yang digunakan dalam penelitian ini yaitu data Twitter berupa *tweet* dan data Gazetteer atau nama tempat. Data Twitter digunakan sebagai data latih pada sistem NER dan sistem klasifikasi kejadian. Sedangkan data Gazetteer digunakan untuk data latih tambahan pada sistem NER.

Twitter API dan *crawling* data

Data Twitter didapatkan dengan cara *crawling* menggunakan Twitter API secara gratis dan bebas. Dengan menggunakan API Twitter pengguna dapat melakukan *crawling* berdasarkan; kata kunci, id pengguna, waktu/tanggal, dan atau lokasi. Pada penelitian ini digunakan pencarian berdasarkan kata kunci, id pengguna, dan atau lokasi pengguna. Pengambilan data dilakukan pada tanggal 3 Maret – 20 Juli 2018.

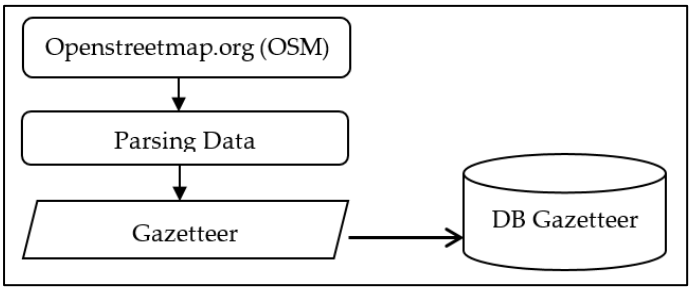
Pencarian berdasarkan kata kunci menggunakan kata seperti; 'kebakaran', 'macet', 'gempa', 'banjir', 'kecelakaan', dan 'longsor'. Sedangkan pencarian berdasarkan *user id* dengan mengikuti *timeline* Twitter dari akun Radio Suara Surabaya (@e100ss), Dishub Surabaya (@sits_dishubsby), dan BMKG (@infoBMKG). Data hasil *crawling* disimpan dalam *file* dengan format txt. Yang nantinya digunakan untuk data latih dan data uji pada sistem. Alur *crawling* data pada Twitter ditunjukkan pada Gambar 2.



Gambar 2. Alur *crawling* data Twitter menggunakan API Twitter

***Gazetteer* dan *parsing* data**

Gazetteer atau nama tempat didapatkan dengan melakukan *parsing* data dari openstreetmap.org (OSM) yang merupakan layanan peta digital. Data seluruh *gazetteer* didapatkan dalam bentuk *file* xml dengan melakukan pembatasan wilayah geografis sekitar kota Surabaya dan Sidoarjo sebagai area lingkup penelitian. Data yang berupa xml diuraikan untuk mendapatkan jenis data yang dibutuhkan seperti; nama kota, nama jalan, dan nama tempat/gedung. Data tersebut kemudian disimpan dalam *database* yang nantinya digunakan sebagai bagian data latih NER. Ilustrasi terkait alur untuk mendapatkan data dan *parsing gazetteer* dapat dilihat pada Gambar 3.



Gambar 3. Alur mendapatkan data dan *parsing gazetteer*

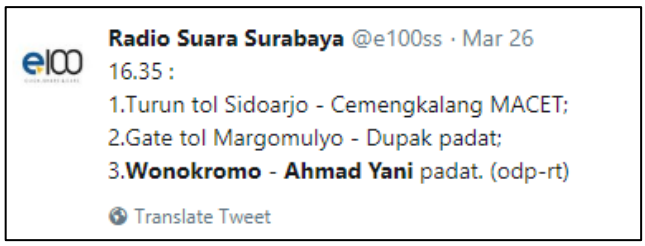
ISSN 2502-3357 (*Online*)
F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90 ISSN 2503-0477 (*Print*)

3.2. Praproses data

Berbagai jenis diterapkan pada penelitian ini guna meningkatkan performa pengenalan entitas maupun maupun klasifikasi kejadian. Praproses yang digunakan dalam sistem ini terdiri dari sembilan jenis praproses antara lain:

a. Menghapus semua tanda baca selain tanda titik (.), strip (-) , dan tanda tanya (?).
b. Mengubah singkatan nama tempat menjadi bentuk panjangnya, seperti; 'TP' menjadi 'Tunjungan Plaza', 'a.yani' menjadi 'Ahmad Yani'.
c. Mengubah singkatan bagian dari nama tempat kebentuk panjangnya seperti; 'per4an' menjadi 'perempatan'.
d. Menghapus ganti baris ('\n') dengan menggantinya menjadi satu spasi.
e. Menghapus kata 'RT' di awal teks *tweet*.
f. Menghapus URL dan *mention*. *Substring* yang sesuai dengan pola URL dan *mention* maka akan dihapus dari teks.
g. Memberi jarak satu spasi antara huruf alfabet dengan tanda baca.
h. Pemenggalan teks *tweet* menjadi beberapa *tweet* jika terdapat tanda pemisah antar *tweet* berupa angka yang diikuti dengan tanda titik. '1. ', '2. ', dan seterusnya.
i. *Casefolding*, mengubah semua huruf menjadi jenis huruf kecil.

Tidak semua tanda baca pada teks *tweet* dihapus dikarenakan tanda baca yang tidak dihapus tersebut mempunyai peran dalam pengenalan entitas. Tanda strip (-) pada teks *tweet* informasi kejadian bisa bermakna sebagai arah seperti pada Gambar 4 dan bisa sebagai fitur dalam entitas *DATE* contoh: '2-Juli-2018'. Begitu juga dengan tanda titik (.), titik berperan sebagai fitur penanda entitas *TIME* contoh '14.35' dan juga sebagai pemisah kalimat yang juga sebagai pemisah, jika tanda titik dihapus maka entitas sebelum dan sesudah *tweet* akan dikenali menjadi satu entitas yaitu LOC bahkan mungkin O (*Other*).



Gambar 4. Contoh satu *tweet* dengan informasi dan jenis kejadian lebih dari satu

Sedangkan praproses ke delapan dilakukan untuk memisahkan beberapa informasi yang mungkin berbeda secara jenis kejadian. Jika *tweet* tidak dipisah maka akan terjadi kesalahan dalam melakukan klasifikasi jenis kejadian karena *tweet* berisi lebih dari satu kejadian. Contoh satu *tweet* dengan informasi dan jenis kejadian lebih dari satu ditunjukkan pada Gambar 4.

Tabel 1. Jenis entitas bernama yang digunakan dan jumlahnya

Label	Contoh	Jumlah
LOC (Location)	Kertajaya, Gubeng, Wonokromo	988
GPE (Geographical Entity)	Surabaya, Malang, Kediri, Blitar	333
BLD (Building)	Taman Pelangi, Royal Plaza	247
NPL (Natural Place)	Gunung Bromo, Sungai Brantas	18
HWYMSE (Highway Measurement)	Km 20 , KM 20.120	37
OBJ (Object)	Truk, Avanza, Motor, Mobil	518
MSE (Measurement)	1 Km, 5.2 SR, 20 cm	230
TIME	12.45, 16:40	193
DATE	12/12/2018, 1-Juli-2018	79
O (*Other*)	Saya, aku, tanya, arah, dimana	11.851
Total		**16.845**

Pengenalan entitas bernama (NeuroNER)

Data latih yang digunakan untuk NeuroNER sebanyak 1.152 *tweet* atau sebanyak 16.845 token. Jenis entitas dalam penelitian ini dibagi menjadi sembilan jenis entitas yaitu LOC, GPE, BLD, NPL,

F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

HWYMSE, OBJ, MSE, *TIME*, dan *DATE*. Penjelasan rinci tentang jenis entitas dijelaskan pada Tabel 1. Untuk format pelabelan jenis entitas data latih pada penelitian ini menggunakan format BIO. B menunjukkan *begin*, atau kata pertama dari entitas, I menunjukkan *Inside* atau kata berikutnya setelah kata pertama dari entitas. Contoh pelabelan data latih ditunjukkan pada Tabel 2. Data yang sudah melalui tahap praproses akan masuk sebagai *input* dari modul *NeuroNER*. Sedangkan hasil keluaran dari NeuroNER nantinya akan digunakan sebagai input dari proses klasifikasi. Contoh keluaran proses ini seperti pada Tabel 3.

Tabel 2. Contoh pelabelan data latih pengenalan entitas bernama

Teks *Tweet*	Token	Label BIO
14.30 ahmad yani arah wonokromo macet	14.30	TIME
	ahmad	B-LOC
	yani	I-LOC
	arah	O
	wonokromo	B-LOC
	macet	O
Ahmad yani depan royal plaza padat merayap antrian hingga 2 km	Ahmad	B-LOC
	yani	I-LOC
	depan	O
	royal	B-BLD
	plaza	I-BLD
	padat	O
	merayap	O
	antrian	O
	hingga	O
	2	B-MSE
	km	I-MSE

Tabel 3. Contoh hasil kelauran proses NeuroNER

Teks hasil praproses	Teks Keluaran NeuroNER
14.30 ahmad yani arah **wonokromo** macet	**TIME LOC** arah **LOC** macet
ahmad yani depan **royal plaza** padat merayap antrian hingga **2 km**	**LOC** depan **BLD** padat merayap antrian hingga **MSE**
tol sumo km 722 arah **surabaya** ada pembakaran lahan	**LOC HWYMSE** arah **GPE** ada pembakaran lahan
15.38: info awal kecelakaan **grandmax** nabrak pembatas jalan di **km 32 tol sidoarjo** arah **porong** .	**TIME** info awal kecelakaan **OBJ** nabrak pembatas jalan di **HWYMSE LOC** arah **LOC** .
lahan rumput terbakar di pinggir jalan tol arah **surabaya - rembang**	lahan rumput terbakar di pinggir jalan tol arah **GPE – GPE**
macet arah **waduk nipah** ada kegiatan warga	macet arah **NPL** ada kegiatan warga
gempa **5.3 sr** guncang **malang**	Gempa **MSE** guncang **GPE**

Klasifikasi RCNN

Data latih yang digunakan sebanyak 814 *tweet*, hasil dari *crawling* pada tahap pengumpulan data. Pada penelitian ini jenis kejadian atau kelas dibagi menjadi empat yaitu: Lalu-Lintas, kebakaran, bencana-Alam, noninformasi kejadian penting. Sedangkan Klasifier yang digunakan untuk melakukan klasifikasi terhadap teks *tweet* adalah algoritma RCNN berdasarkan subbab 2.3.

ISSN 2502-3357 (*Online*)
F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90 ISSN 2503-0477 (*Print*)

Setelah melalui tahap pengenalan entitas bernama maka proses akan diteruskan ke modul klasifikasi jenis kejadian jika *tweet* terdeteksi mempunyai minimal satu entitas bernama yang merepresentasikan lokasi seperti; LOC, GPE, BLD, dan NPL. Maka, input dari RCNN bukan data dari hasil praproses melainkan data keluaran dari proses pengenalan entitas bernama. Contoh pelabelan data latih klasifikasi jenis kejadian ditunjukkan pada Tabel 4. Sedangkan hasil dari uji coba ditunjukkan pada Tabel 5.

Tabel 4. Contoh pelabelan data latih klasifikasi jenis kejadian dan jumlah data *tweet* masing-masing kelas

Teks Data Latih	Label Kelas	Jumlah
LOC arah LOC macet parah	Lalu-Lintas	150
Banjir di LOC mencapai MSE	Bencana-alam	114
Kebakran terjadi di BLD jl LOC, 2 unit pmk meluncur ke lokasi	Kebakaran	122
Mohon info arus dari GPE arah GPE apa macet? thx	Noninformasi kejadian	18
Total		**814**

Tabel 5. Hasil uji coba klasifikasi jenis kejadian

Teks Input dari NeuroNER	Label Kelas	Keterangan
yo iki sing garai macet min . .. podo mlipir kabeh rebutan dalan	Non-Informasi Kejadian	Tidak terdapat entitas lokasi
TIME LOC arah LOC **lancar**	Non-Informasi Kejadian	'Lancar' bukan termasuk informasi kejadian
GPE jadi kota prioritas penggunaan ldi GPE	Non-Informasi Kejadian	Informasi tidak berdampak negatif pada masyarakat
Sblm LOC **macet**	Lalu-Lintas	'Macet' termasuk informasi kejadian yang berdampak negatif pada efisiensi waktu pengguna jalan
TIME info awal **kecelakaan** tunggal OBJ menabrak pembatas jalan di LOC arah LOC sekitar HWYMSE .	Lalu-Lintas	'Kecelakaan' termasuk informasi seperti halnya macet
kebakaran kapal di LOC	Kebakaran	'Kebakaran' berdampak negatif bagi nyawa seseorang juga bisa berakibat kemacetan
GPE di guncang **gempa MSE**	Bencana-Alam	'Gempa' merupakan kejadian yang berdampak negatif bagi nyawa seseorang

4. Hasil dan Pembahasan

Tabel 6. Confusion matrix hasil percobaan untuk masing-masing entitas

Entitas	LOC	GPE	BLD	HWYMSE	NPL	TIME	DATE	MSE	OBJ	Other
LOC	226							1		6
GPE		72							1	
BLD	1		35							
HWYMSE	1			8						
NPL					4					
TIME						65		1		
DATE						1	8	1		
MSE						1	2	10		1
OBJ			1					1	67	
Other	8					1	1	2	2	2000

Percobaan dilakukan dengan menggunakan 391 *tweet* yang sudah didapatkan pada tahap pengumpulan data selain yang digunakan untuk data latih. Uji coba pertama adalah dengan menguji

F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

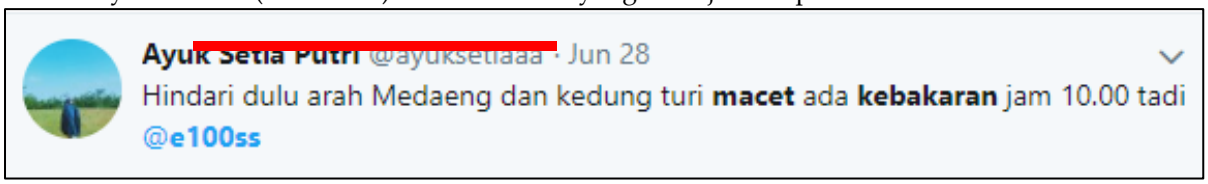
performa dari pengenalan entitas bernama dengan NeuroNER menggunakan perhitungan *precision, recall,* dan *f-measure.* Dari hasil ujicoba yang menunjukkan bahwa NeuroNER bekerja sangat baik dalam mengenali entitas bernama pada teks *tweet* dengan nilai rata-rata *precision, recall,* dan *f-measure* masing masing 96,56%, 95,89%, dan 96,21%. Data hasil percobaan dituangkan dalam tabel *confusion matrix* seperti pada Tabel 6. Dan hasil perhitungan *precision*, *recall*, dan *f-measure* pada Tabel 7.

Tabel 7. Hasil perhitungan *precision, recall,* dan *f-measure* pada hasil uji coba

Entitas	Prec (%)	*Recall* (%)	F1 (%)
LOC	97,00	95,76	96,38
GPE	98,63	100,00	99,31
BLD	97,22	97,22	97,22
HWYMSE	95,24	100,00	97,56
NPL	100,00	100,00	100,00
TIME	98,48	95,59	97,01
DATE	93,10	90,00	91,53
MSE	89,47	85,00	87,18
OBJ	97,10	95,71	96,40
Other	99,30	99,65	99,48
Average	**96,56**	**95,89**	**96,21**

Pada hasil uji coba klasifikasi jenis kejadian, data hasil uji coba dituangkan dalam tabel *confusion matrix* seperti pada Tabel 8. Pada Tabel 8 terlihat jumlah kejadian yang terdetksi melebihi jumlah data uji coba yang hanya berjumlah 391 *tweet*, hal ini dikarenakan satu *tweet* bisa berisi lebih dari satu informasi kejadian yang berbeda seperti pada Gambar 4.

Dari hasil ujicoba didapatkan nilai rata-rata *precision, recall,* dan *f-measure* masing-masing 92,44%, 94,76%, dan 93,53% yang dapat dilihat secara rinci pada Tabel 9. Hal ini menunjukkan bahwa sistem bekerja dengan baik. Meski demikian, masih ada yang diperlu diperbaiki dalam sistem ini yaitu sistem akan tetap mendeteksi atau mengklasifikasikan *tweet* informasi kejadian sebagai satu jenis kejadian meskipun dalam *tweet* terdapat lebih dari satu jenis kejadian. Contoh *tweet* dengan jenis kejadian lebih dari satu yaitu macet (lalu-lintas) dan kebakaran yang ditunjukkan pada Gambar 5.



Gambar 5. Contoh *tweet* dengan jenis kejadian lebih dari satu

Tabel 8. *Confusion matrix* hasil percobaan untuk masing-masing entitas

Kelas	Non-Informasi Kejadian Penting	Lalu - Lintas	Kebakaran	Bencana-Alam
Non-Informasi Kejadian Penting	150	8	2	4
Lalu – Lintas	5	140	2	2
Kebakaran	1		46	
Bencana-Alam	2			45

Tabel 9. Hasil perhitungan *precision, recall,* dan *f-measure* pada hasil uji coba klasifikasi kejadian

Kelas	Prec (%)	*Recall* (%)	F1 (%)
Non-Informasi Kejadian Penting	94,94	91,46	93,17
Lalu - Lintas	94,59	93,96	94,28
Kebakaran	92,00	97,87	94,85
Bencana-Alam	88,24	95,74	91,84
Rata-rata	**92,44**	**94,76**	**93,53**

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

5. Kesimpulan

Penggabungan metode antara NeuroNER dan RCNN yang diusulkan mampu melakukan deteksi atau klasifikasi jenis kejadian dengan baik. Dengan menerapkan pengenalan entitas lokasi sebelum proses klasifikasi sebagai syarat utama dari informasi jenis kejadian penting. Berdasarkan perhitungan *precission, recall,* dan *f-measure* didapatkan nilai rata-rata masing-masing 92,44%, 94,76%, dan 93,53%.

Beberapa kesalahan yang dialami sistem saat uji coba antara lain: 1) Pada pengenalaan entitas bernama, terdapatnya entitas nama lokasi dengan yang unik serta tidak terdapat pada data latih dan *Gazetteer* membuat sistem gagal mengenali kelas entitias dengan benar. Hal ini terjadi karena memang informasi kejadian berada di luar area wilayah penelitian, dan 2) Penggunaan kosa kata baru yang tidak baku untuk menginformasikan kejadian, sehingga sistem gagal mengklasifikasikan informasi dengan benar.

Untuk pengembangan kedepannya, diharapkan sistem dapat melakukan klasfikasi *multilabel* dengan menggunakan klasifier *multilabel* untuk mengatasi masalah *tweet* informasi yang berisi lebih dari satu kejadian. Selain itu, pengenalan entitas bernama khususnya nama tempat/lokasi dipengaruhi oleh banyaknya data nama entitas lokasi yang dilatih dalam sistem NeuroNER, sehingga diperlukan lagi banyak data *Gazetteer* agar dapat diterapkan pada area lingkup penelitian lebih luas lagi mungkin lingkup pulau Jawa atau bahkan nasional.

6. Referensi

Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks, 2*(3), 108-121.

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2017). NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Proceedings of the 2017 EMNLP System Demonstrations* (pp. 97-102). Copenhagen: Association for Computational Linguistics.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica, 17*(4), 635-667.

Gu, Y., Qian, Z. S., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies, 67*, 321-342.

Hasby, M., & Khodra, M. L. (2013). Optimal path finding based on traffic information extraction from Twitter. *International Conference on ICT for Smart Society.* Jakarta: IEEE.

He, K., Li, Y., Soundarajan, S., & Hopcroft, J. E. (2018). Hidden Community Detection in Social Networks. *Information Sciences, 425*(January 2018), 92-106.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CORR*.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access, 5*, 2870-2879.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2267-2273). Austin: AAAI.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *Proceedings of NAACL-HLT 2016* (pp. 260–270). San Diego: Association for Computational Linguistics.

Liu, X., & Zhou, M. (2013). Two-stage NER for tweets with clustering. *Information Processing & Management, 49*(1), 264-273.

Najibullah, A., & Mingyan, W. (2015). Otomatisasi Peringkasan Dokumen Sebagai Pendukung Sistem Manajemen Surat. *Register: Jurnal Ilmiah Teknologi Sistem Informasi, 1*(1), 1-6.

Nidhi, R. H., & Annappa, B. (2017). Twitter-user recommender system using tweets: A content-based approach. *2017 International Conference on Computational Intelligence in Data Science(ICCIDS).* Chennai: IEEE.

Perdana, R. S., Fatichah, C., & Purwitasari, D. (2015). Pemilihan kata kunci untuk deteksi kejadian trivial pada dokumen Twitter menggunakan Autocorrelation Wavelet Coefficients. *JUTI, 13*(2), 152-159.

ISSN 2502-3357 (*Online*)
F. N. Putra & C. Fatichah /Register 4 (2) 2018 81-90 ISSN 2503-0477 (*Print*)

Putra, F. N., Effendi, A., & Arifin, A. Z. (2018). Pembobotan Kata pada Query Expansion dengan Tesaurus dalam Pencarian Dokumen Bahasa Indonesia. *Jurnal Linguistik Komputasional (JLK), 1*(1), 17-22.