TESIS – KI142502

Ekstraksi Informasi Menggunakan Kombinasi Metode *NeuroNER*, *Neural Relation Extraction*, dan FASM pada Deteksi Kejadian dari Data *Stream Twitter*

Fatra Nonggala Putra
5116201057

DOSEN PEMBIMBING
Dr. Eng. Chastine Fatichah, M.Kom

PROGRAM MAGISTER
BIDANG KEAHLIAN KOMPUTASI CERDAS DAN VISI
DEPARTEMEN INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M.Kom.)
di
Institut Teknologi Sepuluh Nopember Surabaya

oleh:
Fatra Nonggala Putra
NRP. 05111650010057

Dengan judul :
Ekstraksi Informasi Menggunakan Kombinasi Metode *NeuroNER*, *Neural Relation Extraction*, dan FASM pada Deteksi Kejadian dari Data Stream Twitter

Tanggal Ujian : 24 Juli 2018
Periode Wisuda : September 2018

Disetujui oleh:

Dr. Eng. Chastine Fatichah, M.Kom.
NIP. 19751220 2001122002 ..... (Pembimbing 1)

Prof. Ir. Handayani Tjandrasa, M.Sc., Ph.D.
NIP. 194908231976032001 ..... (Penguji 1)

Dr. Agus Zainal Arifin, S.Kom., M.Kom.
NIP. 197208091995121001 ..... (Penguji 2)

Dr.Eng. Darlis Herumurti, S.Kom., M.Kom.
NIP. 197712172003121001 ..... (Penguji 3)

Dekan Fakultas Teknologi Informasi dan Komunikasi,

Agus Zainal Arifin, S.Kom., M.Kom.
NIP. 197208091995121001

[*Halaman ini sengaja dikosongkan*]

Ekstraksi Informasi Menggunakan Kombinasi Metode *NeuroNER, Neural Relation Extraction*, dan FASM pada Deteksi Kejadian dari Data Stream Twitter

Nama Mahasiswa	: Fatra Nonggala Putra
NRP	: 5116201057
Pembimbing I	: Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

ABSTRAK

Pemanfaatan twitter untuk deteksi kejadian bencana alam dan lalu-lintas telah dibahas dalam banyak penelitian yang sudah ada. Banyak Informasi yang dibagikan oleh pengguna Twitter dari akun individu maupun akun milik lembaga pemerintahan dan media yang berupa *tweet* informasi kejadian penting yang diperlukan oleh masyarakat. Dengan memanfaatkan API Twitter pengguna bisa mendapatkan data postingan Twitter secara bebas dan gratis berdasarkan kata kunci, id pengguna, dan *geo-location* yang diinginkan.

Dalam penelitian ini digunakan gabungan metode *NeuroNER*, *NeuralRE*, dan FASM untuk deteksi kejadian dengan melakukan ekstraksi informasi pada data *stream Twitter*. Beberapa tahap penelitian dilakukan untuk mendapatkan hasil yang optimal. Pertama, tahap pengambilan data dan prapemrosesan. Kedua, informasi kejadian haruslah memiliki entitas lokasi yang valid. Untuk itu digunakan metode *neuro named entity recognition* (*NeuroNER*) untuk mengenali entitas bernama khususnya entitas lokasi pada data *tweet*. Ketiga, melakukan klasifikasi jenis kejadian kedalam empat kategori kejadian; non-informasi kejadian, bencana alam, lalu-lintas, dan kebakaran dengan menggunakan algoritma klasifikasi *recurrent convolutional neural network* (RCNN). Keempat, dilakukan proses ekstraksi relasi dengan *NeuralRE* untuk mendapatkan relasi antar entitas bernama. Kelima, standarisasi nama lokasi, *geocoding*, dan visualisasi data ke dalam peta digital.

Penelitian menguji gabungan metode yang disusulkan secara parsial maupun keseleruhan. Gabungan metode yang diusulkan bekerja dengan baik untuk melalukan ekstraksi informasi mulai dari tahap *streaming* data hingga visualisasi data. Hal ini ditunjukkan dengan nilai rata-rata perhitungan *precision, recall*, dan *f-measure* secara keseluruhan masing-masing 94,28%, 94,16%, dan 94,22%.

Kata kunci : deteksi kejadian, ekstraksi informasi, *NeuroNER, neural relation extraction, fast approximate string matching* .

Information Extraction Using Combination of NeuroNER, Relation Extraction, and FASM Method on Incidents Detection from Twitter Data Stream

Name	: Fatra Nonggala Putra
Student Identity Number	: 5116201057
Supervisor	: Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

ABSTRACT

Utilization of twitter for the detection of natural disaster and traffic incidents has been discussed in many existing studies. Lots of Information about important incidents shared by Twitter users from personal, government agencies', and media account are useful for the community. By utilizing Twitter API, users can get Twitter data for free based on keywords, user ids, or geo-locations as desired. We proposed combination of NeuroNER, NeuralRE, and FASM as incident detection method by extracting information from Twitter data stream. Our proposed method consists of five main steps. The first step is data retrieval and preprocessing. The second step is entity location recognition using Neuro Named Entity Recognition (NeuroNER) method to detect valid location of the incidents. The third step is event type classification using Recurrent Classification Algorithms Convolutional Neural Network (RCNN). The events are classified into four categories: non-information, natural disasters, traffic, and fire. The fourth step is entity relations extraction process using NeuralRE to identify relationships between named entities. The final step is standardization of the location name, geocoding, and data visualization onto digital map.

This study examines combinations of some steps and the entire steps of proposed method. The proposed method works well enough in extracting information from streaming data step to data visualization with the average value of precision, recall, and f-measure 94,28%, 94,16%, and 94,22% respectively.

Keywords : *incidents detection, information extraction, NeuroNER, neural relation extraction, fast approximate string matching.*

[*Halaman ini sengaja dikosongkan*]

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan kasih dan sayang-Nya kepada kita, sehingga penulis bisa menyelesaikan tesis dengan tepat waktu, dengan judul "**Ekstraksi Informasi Menggunakan Gabungan Metode *Neuro Named Entity Recognition*, *Neural Relation Extraction*, dan FASM pada Deteksi Kejadian dari Data Stream Twitter**" Tujuan dari penyusunan tesis ini guna memenuhi salah satu syarat untuk bisa menempuh ujian magister komputer di Fakultas Teknologi Informasi dan Komunikasi (FTIK) Program Studi S2 Informatika Institut Teknologi Sepuluh Nopember Surabaya (ITS).

Didalam pengerjaan tesis ini telah melibatkan banyak pihak yang sangat membantu dalam banyak hal. Oleh sebab itu, disini penulis sampaikan rasa terima kasih sedalam-dalamnya kepada:

1. Kedua Orang Tua dan Kakak yang selalu memberikan dukungan moril maupun materiil.
2. Ibu Chastine Fatichah, selaku dosen pembimbing yang telah memberikan ilmu dan waktunya kepada penulis untuk menyelesaikan tesis ini.
3. Ibu Handayani Tjandrasa, Bapak Agus Zainal Arifin, dan Bapak Darlis Heru Murti selaku dewan penguji pada sidang tesis peneliti yang telah memberikan saran untuk perbaikan tesis ini.
4. Teman-Teman S2 Informatika 2016 yang tidak dapat disebutkan satu-persatu. Terima kasih atas persahabatan selama masa studi.

Penulis menyadari bahwa tesis ini masih jauh dari sempurna, karena kesempurnaan hanya milik Allah SWT. Oleh karenanya, kritik dan saran guna perbaikan penelitian selatjutnya. Semoga tesis ini bisa bermanfaatkan minimal dapat jadi rujukan perihal kekurangan yang ada di dalamnya.

Surabaya, 31 Juli 2018

Fatra Nonggala Putra

[*Halaman ini sengaja dikosongkan*]

DAFTAR ISI

DAFTAR GAMBAR

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

BAB 1
PENDAHULUAN

Pada Bab ini akan dijelaskan mengenai beberapa hal dasar dalam pembuatan proposal penelitian yang meliputi latar belakang, perumusan masalah, tujuan, manfaat, kontribusi penelitian, dan batasan masalah.

1.1. Latar Belakang

Berkembang pesatnya dunia teknologi informasi memudahkan manusia untuk bertukar informasi dengan cepat dan mudah. Salah satu media pertukaran informasi saat ini adalah microblogging seperti Twitter. Dengan menghasilkan sekitar 500 juta tweet setiap harinya dan pengguna aktif setiap bulannya mencapai 330 juta pengguna, maka tidak berlebihan jika twitter layak disebut microblogging terpopuler saat ini. Banyak pengguna memanfaatkan twitter sebagai sarana berbagi informasi kejadian penting, baik pengguna individu maupun pengguna sebagai representatif lembaga atau perusahaan yang disebut sebagai *User Influence* (UI). Contoh dari UI yang menggunakan Twitter sebagai sarana berbagi informasi kejadian seperti BMKG, Dishub Surabaya, dan Radio Suara Surabaya.

Sejumlah penelitian telah dilakukan guna memanfaatkan masifnya data tweet yang diposting oleh pengguna twitter. Mulai dari analisa pengaruh penggunaan twitter untuk kampanye politik terhadap hasil pemilihan [1], deteksi komunitas tesembunyi pada jaringan pertemanan pengguna Twitter [2], analisa sentimen tweet terhadap topik tertentu [3][4], analisa komunikasi politik dalam Twitter [5], Twitter sebagai alat komunikasi dan manajemen bahaya [6], integrasi sosial medial dengan aplikasi sistem deteksi bahaya [7],dan deteksi kejadian alam dan lalu-lintas [8][9]. Dalam penelitian ini akan fokus membahas tentang deteksi kejadian alam dan lalu-lintas menggunakan data *tweet* dari Twitter.

Beberapa penelitian yang menggunakan data twitter untuk mendeteksi kejadian secara (mendekati) waktu nyata telah dilakukan antara lain oleh Yiming Gu dkk [8] dengan melakukan pengklasifikasian biner menggunakan semi-Naive Bayes terhadap data stream twitter yang masuk untuk dikenali sebagai tweet informatif (macet, kecelakaan, kebakaran, gempa, dll) dan tweet non-informatif

(selain tweet kategori tweet informatif) kemudian dilakukan ekstraksi geo-lokasi dan klasifikasi jenis kejadian menjadi 5 kategori menggunakan algoritma klasifikasi sLDA. Penelitian lainnya [9] melakukan ekstraksi lokasi kemacetan menggunakan *named-entity-recognition* (NER) dan *relation extraction* (RE) yang kemudian digunakan untuk mencari jalur perjalanan lain yang paling optimal.

Kejadian penting seperti kecelakaan, kebakaran, kemacetan atau insiden lainnya yang dapat menimbulkan dampak negatif kepada masyarakat perlu diketahui sedini mungkin oleh masyarakat umum agar dapat terhindar dari dampak kejadian. Dan untuk pihak berwajib, mengetahui informasi kejadian penting sedini mungkin merupakan suatu keharusan agar dapat segera melakukan tindakan untuk mengatasi masalah tersebut. Untuk itu, pihak berwajib perlu mempunyai database setiap kejadian secara sistematis dan terstruktur seperti jenis, waktu, dan lokasi kejadian, agar dapat mengetahui apakah suatu kejadian itu termasuk kejadian berulang atau tidak berulang sehingga dapat dilakukan analisa dan tindakan untuk mengatasi permasalahan yang ada.

Tweet atau postingan pengguna twitter merupakan data teks yang tidak terstruktur. Kebanyakan sebuah tweet mengandung akronim dan kata tidak baku khas percakapan media sosial sehingga sulit untuk dilakukan ekstraksi informasi secara otomatis. Untuk melakukan ekstraksi informasi pada data twitter memerlukan tahapan-tahapan praproses khusus agar data twitter yang memiliki keunikan dalam penulisan dapat diolah menjadi data yang baik dan dapat dimengerti sistem. Data yang baik memudahkan sistem untuk melakukan klasifikasi dan ekstraksi informasi sehingga mendapatkan hasil yang lebih baik.

Beberapa masalah khusus pada data twitter untuk proses deteksi kejadian antara lain; 1) penggunaan akronim yang tidak baku contoh: ‘sampai’ ditulis ‘smp’, 2) penulisan kata agar lebih singkat dengan tambahan angka contoh: ‘pertigaan’ ditulis ‘per3an’, 3) penulisan kata dengan tambahan huruf yang berulang contoh: ‘macet’ ditulis ‘maaceeet’, 4) penggunaan karakter ‘-‘ sebagai pengganti kata sampai, antara, dan sekitar contoh: surabaya – sidoarjo, dan 5) kesalahan penulisan nama tempat yang merupakan informasi penting selain informasi jenis kejadian contoh: ‘jalan mayjend sungkono’ ditulis ‘jalan my jen sungkono’.

Pada fase praproses, untuk menghasilkan data yang baik diperlukan metode yang tepat untuk mengatasi masalah khusus pada data teks Twitter yang tidak sama dengan data teks lainnya seperti dokumen berita dan artikel ilmiah. Beberapa jenis teknik praproses memiliki pengaruh yang berbeda dalam performa sistem klasifikasi. Seperti dalam [4], teknik penghapusan URL dan *stop words* ternyata tidak berdampak dalam meningkatkan performa klasifikasi namun, memiliki dampak dalam performa sistem yaitu mengurangi ukuran dari fitur model. Kemudian teknik mengganti akronim dan singkatan tidak baku menjadi bentuk umum memiliki dampak yang signifikan dalam peningkatkan performa Sensitivity klasifikasi. Sedangkan penggunaan *stemming/lemma* justru menurunkan performa akurasi klasifikasi.

Berikutnya fase pengenalan entitas bernama (NER) untuk mengenali entitas bernama seperti nama lokasi dan objek. NER berperan untuk mengenali kata atau frase sebagai suatu entitas tertentu. Jika hasil ekstraksi informasi tidak dapat mengenali entitas lokasi secara benar maka akan menjadikan informasi tersebut menjadi tidak valid dan kurang berguna. Penggunaan model baru dari NER yaitu *NeuroNER* [7][8] pada dokumen online untuk melakukan pengenalan entitas khusus seperti nama orang, lokasi, dan organisasi menghasilkan performa yang lebih baik daripada model NER sebelumnya [10]. Pada umumnya, penggunaan NER tidak akan berjalan baik untuk ekstraksi entitas bernama jika teks nama entitas ditulis dengan ejaan yang salah atau tidak sesuai dengan pola data latih yang digunakan pada fase pelatihan. Misalnya, pola umum penulisan nama tempat dalam data latih dokumen formal adalah menggunakan huruf kapital diawal kata, maka NER akan memiliki persentase keberhasilan kecil untuk dapat mendeteksi entitas lokasi dengan pola penulisan huruf kecil sebagai huruf pertama pada data uji. Sedangkan penulisan teks pada twitter umumnya ditulis menggunakan huruf kecil meskipun kata tersebut merupakan nama entitas lokasi.

Pengenalan terhadap entitas bernama saja tidak cukup untuk mendapatkan pengetahuan atau informasi dari suatu tweet. Pada informasi kemacetan di jalan raya umumnya memberikan informasi tentang arah dari kemacetan karena suatu jalur umumnya memiliki arah (menuju a atau menuju b), sedangkan a ke b dan b ke a merupakan entitas yang berbeda dalam jalur transportasi. Sehingga diperlukan

pengenalan arah untuk mengetahui jalur mana yang sedang terjadi kemacetan. Permasalahan memahami hubungan rangkaian lokasi yang saling berhubungan dapat dipecahkan dengan penggunaan *relation extraction*.

Relation Extraction (RE) pada dasarnya sama dengan konsep klasifikasi teks biasa . Beberapa penelitian terbaru menggunakan model *neural network* untuk melakukan klasifikasi relasi pada teks [12][13][[14]. RE model *neural netwok* lebih populer disebut *Neural Relation Extraction* (*NeuralRE*). *NeuralRE* memilki kemampuan yang lebih baik daripada metode ekstraksi relasi konvensional.

Oleh karena itu, penelitian ini berfokus pada ekstraksi informasi kejadian pada data stream Twitter untuk mengatasi permasalahan ekstraksi informasi dengan menggunakan kombinasi metode *NeuroNER*, *NeuralRE*, dan FASM. *Neural Relation Extration* (*NeuralRE*) digunakan untuk melakukan ekstraksi relasi antar entitas lokasi jika terdapat lebih dari satu entitas. Entitas lokasi yang berhasil dikenali oleh *NeuroNER* kemungkinan ditulis tidak sesuai dengan standar penulisan nama lokasi pada Gazetteer. Maka dari itu diperlukan proses standarisasi nama, metode yang digunakan untuk proses standarisasi nama lokasi adalah *Fast Aproximate String Matching* (*FASM*) yang merupakan varian *aproximate string matching* yang memanfaatkan teknik *edit distance Damerau-Levenshtein Distance* (DLD) untuk mendapatkan *string* dengan kemiripan paling tinggi dengan performa kecepatan yang tinggi [15]. Standarisasi nama memudahkan proses *geocoding* untuk visualisasi informasi kejadian kedalam peta digital.

1.2. Perumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah sebagai berikut.

1. Bagaimana cara melakukan praproses untuk data Twitter pada sistem ekstraksi informasi kejadian?
2. Bagaimana cara melakukan klasifikasi tweet kejadian penting atau non-kejadian penting menggunakan klasifier RCNN?
3. Bagaimana cara melakukan ekstraksi informasi pada data Twitter menggunakan metode *NeuroNER*?
4. Bagaimana mengetahui relasi antar entitas menggunakan metode *NeuralRE*?

5. Bagaimana cara standarisasi entitas nama lokasi menggunakan metode FASM?
6. Bagaimana melakukan visualisasi pemetaan data informasi kejadian pada peta digital berdasarkan lokasi dan jenis kejadian?

1.3. Tujuan

Tujuan yang akan dicapai dalam pembuatan tesis ini adalah ekstraksi informasi menggunakan kombinasi metode *NeuroNER*, *Neural Relation Extraction*, dan FASM pada Deteksi Kejadian dari Data *Stream* Twitter.

1.4. Manfaat

Manfaat dari penelitian ini adalah:

a. Mengetahui jenis praproses yang sesuai untuk data *tweet* pada ekstraksi informasi kejadian
b. Menghasilkan metodologi terbaik untuk melakukan ekstraksi informasi pada data stream Twitter.
c. Memberikan informasi secara cepat dan utuh kepada masyarakat dengan melakukan pemetaan dan pengelompokan informasi kejadian berdasarkan lokasi dan jenis kejadian secara akurat pada peta digital.
d. Ekstraksi informasi secara sistematis dan tersimpan dalam sistem dapat digunakan pihak terkait untuk mengetahui statistik riwayat dari suatu tempat terhadap suatu kejadian yang nantinya dapat dikategorikan menjadi kejadian berulang dan tidak berulang untuk mendapatkan antisipasi lebih lanjut berdasarkan riwayat yang tersimpan.

1.5. Kontribusi Penelitian

Kontribusi pada penelitian ini adalah penggabungan beberapa metode yaitu *NeuroNER*, *NeuralRE*, dan FASM untuk ekstraksi informasi pada deteksi kejadian dari data stream Twitter. Fokus utama ekstraksi informasi adalah untuk mendapatkan informasi lokasi kejadian dengan menggunakan *NeuroNER* dan mendapatkan relasi antar entitas lokasi menggunakan *NeuralRE* serta penggunaan

metode *FASM* untuk standarisasi entitas nama lokasi yang telah dikenali oleh *NeuroNER* dan mendapatkan hubungan antar relasi dengan representasi graf.

1.6. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Wilayah yang dijadikan objek penelitian adalah kota Surabaya dan Sidoarjo
2. Proses pengumpulan dataset dilakukan dengan cara melakukan *crawling* data pada akun twitter Suara Surabaya (@e100ss), Badan Meteorologi, Klimatologi, dan Geofisika (@bmkg), dan DISHUB Surabaya (@sits_dishubsby) pada bulan April dan Mei.
3. Kamus kata khusus pada Twitter bahasa indonesia diekstrak dari hasil pengumpulan tweet pada no.2 dan dilakukan pembuatan persamaan kata bakunya secara manual
4. Gazetteer yang digunakan di-ekstrak dari openstreetmap.org untuk wilayah kota Surabaya dan Sidoarjo.
5. Word2vec yang digunakan dibangun dari seluruh data artikel wikipedia bahasa indonesia.

BAB 2
KAJIAN PUSTAKA

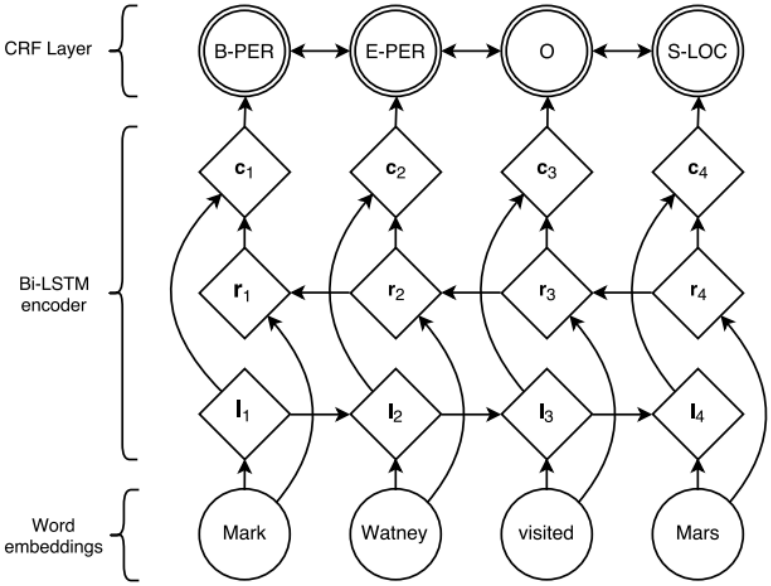
Pada bab ini akan dijelaskan tentang pustaka yang terkait dengan landasan penelitian. Pustaka yang terkait adalah seputar deteksi kejadian, ekstraksi informasi, dan klasifikasi.

2.1 Ekstraksi Informasi

Ekstraksi informasi adalah teknik untuk mendapatkan informasi secara spesifik dari sekumpulan data teks dengan mengubah data teks tidak terstruktur menjadi terstruktur sesuai format yang diinginkan. Ekstraksi informasi biasanya digunakan untuk mendapatkan kata dan frase khusus dalam teks yang disebut sebagai entitas seperti nama lokasi, nama orang, dan nama organiasi. Salah satu teknik ekstraksi informasi adalah *named entity recognition* (NER) [16]]. NER biasanya digunakan pada berbagai domain keilmuan yang berbeda untuk mengenali entitas tertentu. Dalam dunia medis, NER digunakan untuk mengenali entitas nama penyakit pada dokumen medis [17][18]. Dalam deteksi kejadian, NER digunakan untuk mengenali entitas nama lokasi dan tempat [9].

2.2 Neuro Named Entity Recognition (*NeuroNER*)

NeuroNER merupakan modifikasi dari teknik NER sebelumnya yang menggunakan algoritma *Conditional Random Field* (CRF) dengan menambahkan algoritma *Recurrent Neural Network* (RNN) pada model NER. Jenis RNN yang digunakan pada metode ini adalalah *bidirectional Long Short-term Memory Networks* (biLSTM) [19][18]. *NeuroNER* membuat alur prediksi anotasi berjalan lebih baik dengan ANNs dan CRF daripada hanya menggunakan CRF [10]. Arsitektur dari *NeuroNER* ditunjukkan pada Gambar 2.1.



Gambar 2.1 Arsitektur *NeuroNER* [7]

2.3 *Neural Relation Extraction* (NeuralRE)

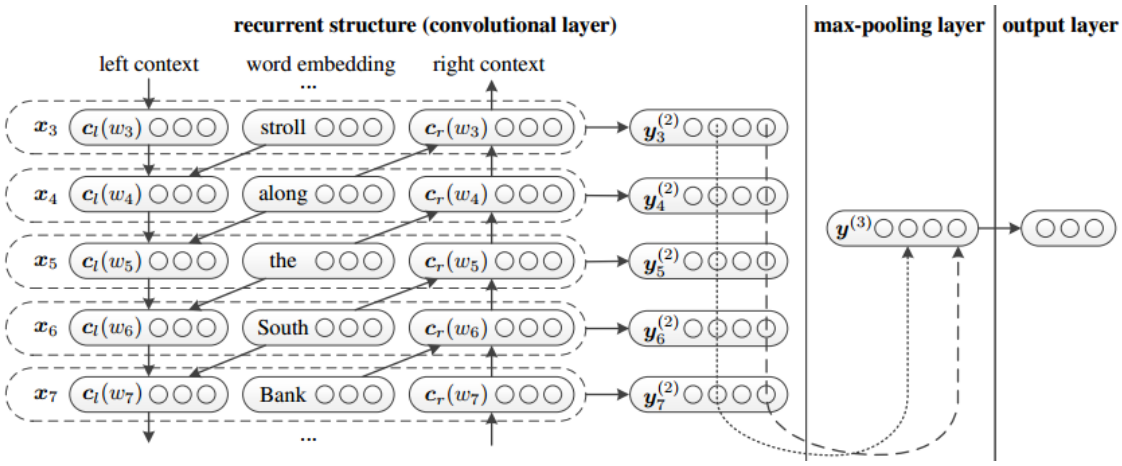
NeuralRE adalah metode yang bertujuan melakukan ekstraksi informasi berupa relasi yang digunakan untuk mengetahui hubungan antar entitas pada teks kalimat [12][13]. RE konvensional seperti pada [9] dianggap kurang efektif karena menggunakan fitur 'buatan tangan' [18]. Model baru dari RE dengan menggunakan model *neural network* yaitu algoritma *Recurrent Convolutional Neural Network* (RCNN) [20]. Berbeda dengan klasifikasi, penyebutan kelas atau kategori hasil proses ekstraksi pada RE disebut dengan relasi. Namun, Prinsip dasar dari *relation extraction* adalah sama seperti prinsip klasifikasi teks yaitu sama-sama melakukan proses pengenalan kelas/relasi dari suatu teks.

2.4 Recurrent Convolutional Neural Network (RCNN) Classification

Recurrent Convolutional Neural Network (RCNN) merupakan algoritma pengembangan dari RNN dan CNN dengan melihat kelemahan dan memanfaatkan kelebihan dari keduanya dalam memecahkan masalah klasifikasi data teks [20][12].

Algoritma klasifikasi RCNN menggunakan *bi-directional recurrent* struktur yang diharapkan dapat mengenalkan sangat sedikit *noise* fitur dibandingkan model *neural network* tradisional berbasis *window*. Selain itu,

digunakan *max-pooling layer* yang secara otomatis menilai semua fitur dan menentukan fitur terbaik yang berperan sebagai kunci dalam klasifikasi teks.



Gambar 2.2 Struktur dari RCNN

2.5 Fast Approximate String Matching

Fast approximate string matching (FASM) dengan memanfaatkan metode *edit distance Damerau-Levenshtein Distance* merupakan modifikasi dari teknik pencarian lainnya yaitu *fuzzy string matching* (FSM) yang memanfaatkan teknik *edit distance Levenshtein Distance* untuk mencari string yang cocok dengan prediksi kemungkinan yang paling mendekati (bukan tepatnya atau pasti) [15]. FASM memiliki performa prediksi dan waktu komputasi yang lebih baik daripada *fuzzy string matching* standar. Contoh dari FASM ditunjukkan dalam tabel 2.1.

Tabel 2.1 Contoh Pencarian Perbaikan String dengan FASM [19]

String	Best Correction	Edit Dist.	Max. Edit Dist.	ms per 1000 lookups
zacamodation	accommodation	4	4	727
acamodation	accommodation	3	3	180
acomodation	accommodation	2	2	33
hous	hous	1	1	24
house	house	0	1	1

2.6 Directed Graf

Graf atau jaringan adalah kumpulan node (V) bersama yang dihubungkan dengan *edges* (E) untuk menghubungan antar node [21]. Suatu graf mempunyai kumpulan node (V) yang terdiri dari {v_1,v_2,v_3, ..., v_n} dan *edges* (E) yang

menghubungkan antar node dalam V, dimana E = {$e_1(v_1,v_2)$, $e_2(v_2,v_3)$, $e_3(v_1,v_3)$, e_n}. Beberapa jenis graf antara lain: 1) *directed-graph*, 2) *undirected-graph*, dan 3) *multi-graph*.

Beberapa istilah khusus pada directed graf antara lain : 1) **Successor**, yaitu node x dihubungkan langsung ke node y (x, y) dan y merupakan successor dari x, 2) **Predecessor**, node x dihubungkan langsung ke node y (x, y) dan x adalah predecessor dari y. Ilustrasi dari successor dan predecessor ditunjukkan pada Gambar 2.3.

Gambar 2.3 Contoh Directed Graf dengan 1 Successor dan 1 Predecessor

2.7 Penelitian yang Relavan

Berdasarkan tabel 2.4 di bawah, maka penulis akan mengadopsi strategi dari G.Yiming dalam memfilter data stream tweet yang masuk dengan melakukan klasifikasi kedalam tweet kejadian penting atau tweet non-kejadian penting. Jika terdeteksi sebagai tweet kejadian penting maka tweet akan masuk ke proses berikutnya. Model NER yang digunakan pada penelitian ini mengadopsi metode *NeuroNER* [10][11] dengan *biLSTM* dan CRF yang bekerja lebih baik daripada metode NER dengan CRF saja dalam mengenali entitas bernama.

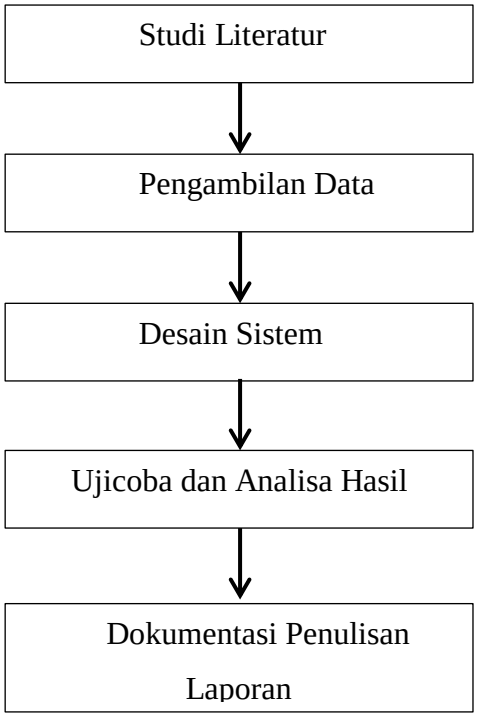
Pada beberapa kasus sering terjadi kesalahan penulisan tweet terutama penulisan nama lokasi. Jika terjadi kesalahan penulisan maka sistem tidak dapat melakukan geocoding untuk memetakan lokasi kejadian. Maka diperlukan standarisasi nama lokasi berdasarkan Gazetteer. Untuk melakukan standarisasi nama lokasi atau alamat digunakan *Fast Approximate String Matching* (FASM) yang bertugas mencari penulisan nama yang benar di dalam Gazetteer.

Tabel 2.2 Penelitian yang Relevan

No	Referensi	Deskripsi	Kelebihan	Kekurangan
1	(G. Yiming, 2016)	- Melakukan filter stream tweet untuk mendeteksi TI dan NTI - Ekstraksi lokasi dengan 3 metode 1) ArcGis Geo-parsing, 2) Hwy Geoparsing, dan 3) Fuzzy Geo-parsing - Klasifikasi jenis kejadian mengguna-kan sLDA	- Memilah tweet yang akan di-proses dengan klasifikasi Naive Bayes - Membedakan lokasi kejadian di jalan umum dan jalan tol dengan template khusus	- Bergantung pada software berbayar arcGIS untuk mengesktrak geocode lokasi. - Jika terjadi kesalahan penulisan lokasi maka sistem tidak dapat melakukan geocode.
2	(J Gelr*ner*, 2013) An algorithm for local geoparsing of microtext	- Metode untuk melakukan geo-parsing pada data *microtext* seperti Twitter dengan menggunakan algoritma LDA	Merumuskan strategi untuk: - Identifikasi singkatan - Identifikasi akronim - Memilih teks ter-baik dari dis-ambiguitas	- Projek masih dalam proses - Penggunaan 3 metode yang berbeda akan memperbesar cost waktu komputasi terutama metode pencocokan nama lokasi dengan Gazetter meng-gunakan Fuzzy
3	(F. Dernoncourt, 2017) *NeuroNER*	- penggunaan ANN dan CRF untuk mengenali entitas bernama seperti nama orang, tempat, dan organisasi.	- Hasil pengenalan-an entitas lebih baik daripada model NER CRF	

BAB 3
METODOLOGI PENELITIAN

Bab ini akan memaparkan tentang metodologi penelitian yang digunakan pada penelitian ini, yang terdiri dari (1) studi literatur, (2) pengambilan data, (3) desain sistem, (4) ujicoba dan analisa hasil, (5) dokumentasi penulisan laporan. Ilustrasi diagram alur metodologi penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram Alur Metodologi Penelitian

Penjelasan tahapan metode penelitian pada Gambar 3.1 akan diterangkan secara terperinci pada subbab berikut.

3.1 Studi Literatur

Tahapan awal dari penelitian dilakukan dengan melakukan kajian berbagai literatur yang berkaitan dengan topik penelitian ekstraksi informasi khususnya pada platform microblogging twitter. Referensi yang digunakan dalam penelitian ini bersumber dari jurnal, konferensi, dan buku yang berkaitan dengan pemrosesan teks khususnya ekstraksi informasi, NER, RE, dan algoritma klasifikasi. Berdasarkan studi literatur yang telah dilakukan, dapat diambil informasi sebagai berikut:

a. Deteksi kejadian lalu-lintas dan bencana alam menggunakan data twitter menjadi lebih murah daripada deteksi kejadian konvensional.
b. Kendala dalam ekstraksi informasi pada sebuah tweet pada umumnya adalah penulisan kata yang tidak baku, terlalu banyak singkatan, dan akronim yang sulit diidentifikasi persamaan kata bakunya.
c. Peningkatan performa ekstraksi informasi pada data tweet masih menjadi tantangan yang perlu diteliti lebih lanjut.
d. Performa ekstraksi informasi kejadian menggunakan NER bergantung pada algoritma dan data latih yang digunakan

3.2 Informasi Kejadian Penting

Istilah informasi kejadian penting terdiri dari 3 kata yaitu informasi, kejadian, dan penting. Setiap kata memiliki definisi yang berbeda. Berdasarkan kamus besar bahasa indonesia daring definisi dari informasi adalah pemberitahuan, atau kabar atau berita tentang sesuatu. Kejadian adalah peristiwa; sesuatu yang terjadi. Sedangkan penting memiliki arti sangat berharga (berguna). Jika digabungkan definisi dari ketiga kata tersebut maka dapat memberikan definisi umum dari informasi kejadian penting sebagai kabar atau berita yang sangat berharga tentang suatu peristiwa yang terjadi.

Berdasarkan [8] dan definisi di atas, maka dapat diformulasikan kembali karakteristik dari suatu informasi dapat disebut sebagai informasi kejadian penting adalah memiliki tiga unsur: 1) **waktu:** informasi berupa kejadian penting dengan waktu pemberitaan mendekati waktu-nyata (*real-tim*e). Jadi, selisih waktu kejadian dengan waktu pemberitaan (*posting*) hampir tidak ada selisih waktu atau dikatakan mendekati waktu-nyata, 2) **dampak:** kejadian yang diinformasikan memiliki dampak negatif atau merugikan masyarakat dalam beraktifitas yang menyangkut keselamatan nyawa maupun kerugian waktu masyarakat yang berada pada sekitar tempat kejadian, 3) **lokasi:** informasi kejadian terbaru yang berpotensi memiliki dampak negatif bagi masyarakat dalam beraktifitas haruslah memiliki entitas lokasi. Tanpa adanya entitas lokasi di dalam informasi tersebut, maka informasi menjadi tidak penting.

3.3 Sistem Deteksi Kejadian

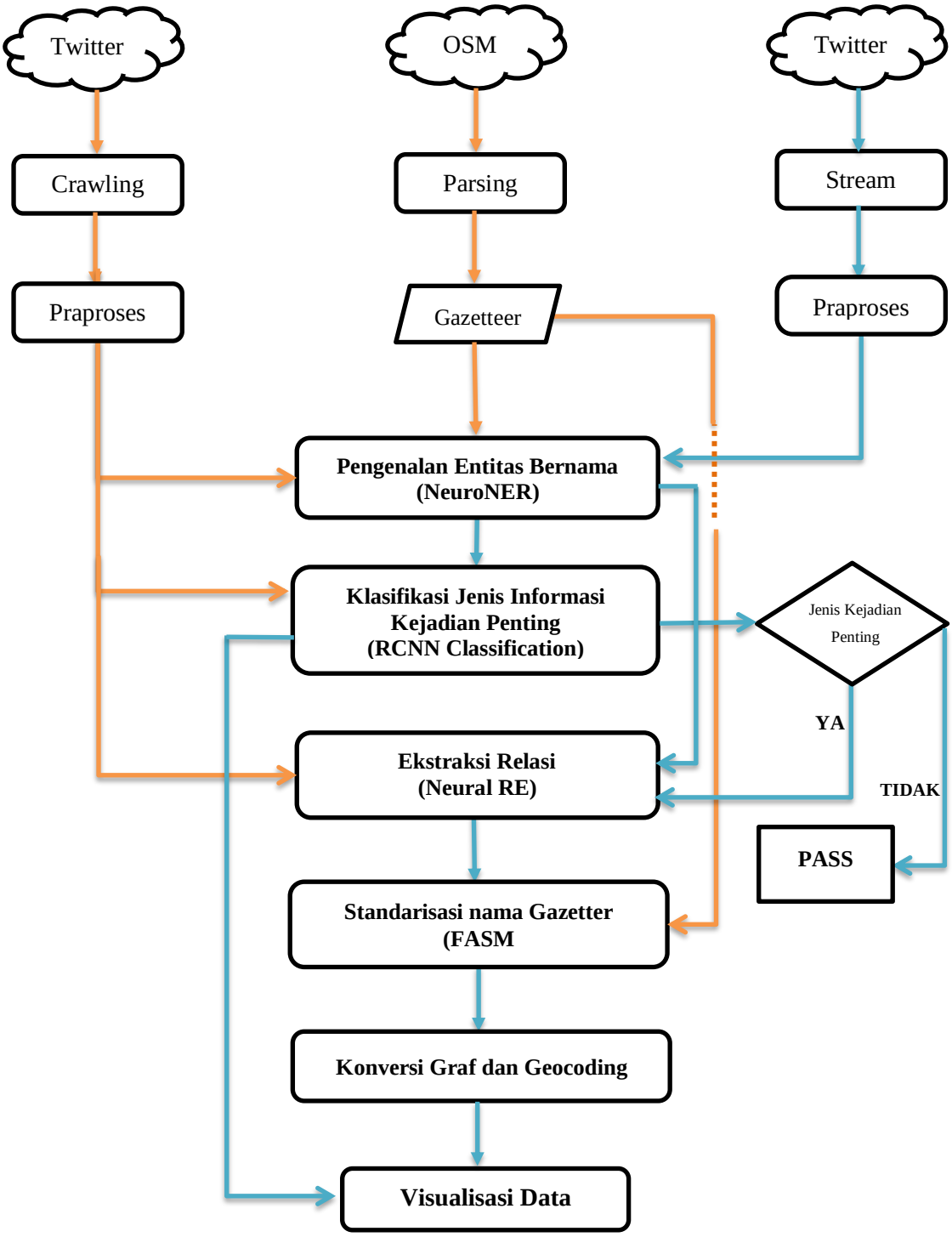
Deteksi kejadian diawali dengan melakukan streaming data pada akun twitter @e100ss dan @sits_dishubsby. Setiap ada data tweet baru akan digunakan sebagai data input yang kemudian dilakukan tahapan praproses untuk membuat data menjadi layak. Dari tahapan praproses data dilakukan klasifikasi untuk mendeteksi tweet yang masuk kelas tweet kejadian penting atau bukan. Jika tweet teridentifikasi sebagai tweet kejadian penting maka akan dilakukan proses selanjutnya seperti pada Gambar 3.2. Gambar 3.2 merupakan diagram alur sistem secara keseluruhan.

3.4 Pengambilan Data

Ada dua jenis data yang dibutuhkan dalam penelitian ini yaitu 1) data Tweet, dan 2) data Gazetteer kota Surabaya dan Sidoarjo. Pada setiap jenis data memerlukan tahapan praproses khusus untuk memastikan data yang akan diolah berupa data yang dapat digunakan.

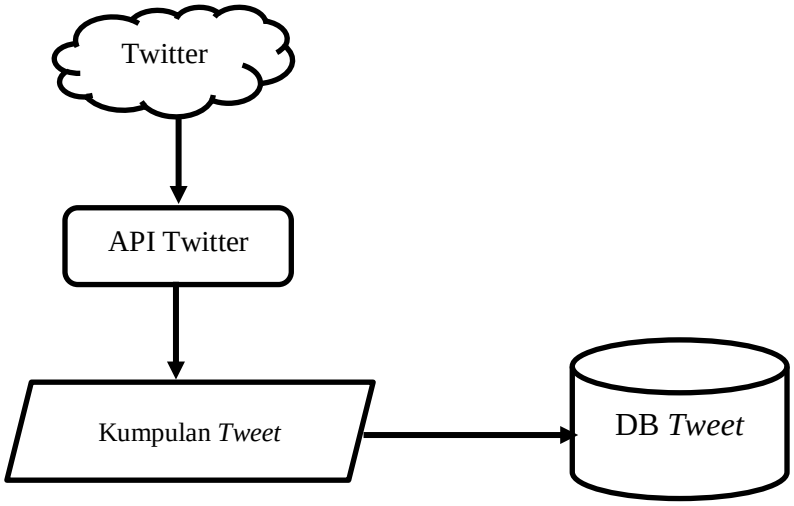
3.4.1 Data Tweet

Pertama, Pengumpulan data tweet dilakukan dengan melakukan *crawling* data *tweet* menggunakan API Twitter yang ditujukan untuk akun Radio Suara Surabaya (@e100ss) dan *crawling* berdasarkan kata kunci yang dilakukan pada bulan Maret-Juni. Diagram alur pengambilan data *tweet* ditunjukkan pada Gambar 3.3.



Gambar 3.2 Desain Sistem Deteksi Kejadian

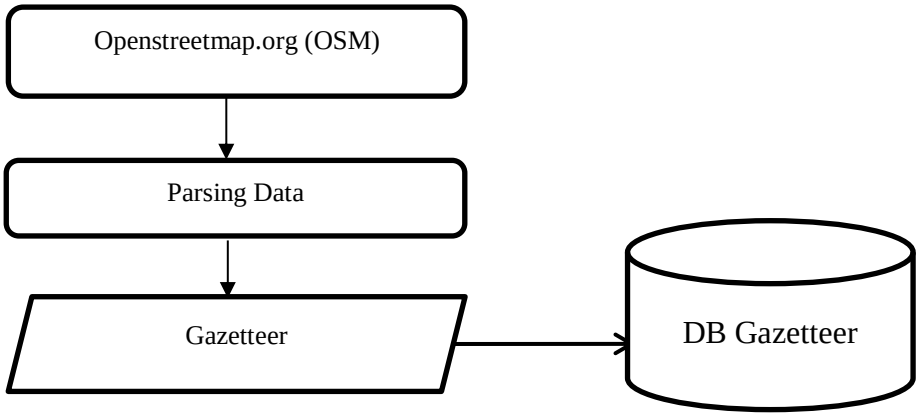
: (Fase *Training*) : (Fase *Testing*)



Gambar 3.3 Desain Sistem Deteksi Kejadian

3.4.2 Data Gazetter

Kedua, mendapatkan Gazetteer di area kota Surabaya dan Sidoarjo. Gazetteer didapatkan dari openstreetmap.org (OSM) dengan melakukan pembatasan wilayah yang akan diambil kemudian hasil file berupa XML dari OSM dilakukan ekstraksi menggunakan bantuan library xmltree python untuk mengekstrak 1) id lokasi, 2) kota atau kawasan, 3) alamat lokasi, 4) nama lokasi, 5) latitude dan longtitude. Alur untuk mendapatkan Gazetteer ditunjukkan pada gambar 3.4 dan contoh dari Gazetter seperti pada Tabel 3.2.



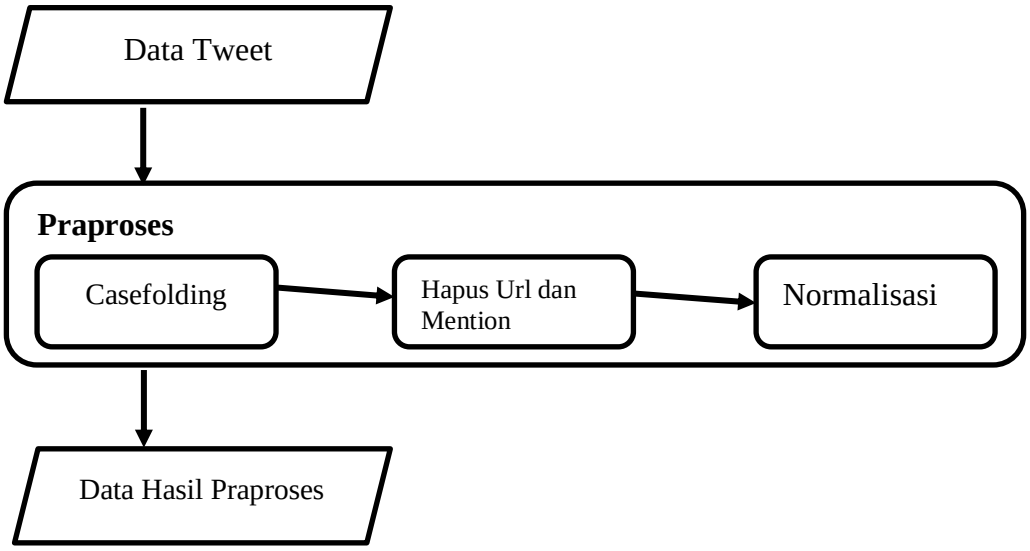
Gambar 3.4 Diagram Alur Pembentukan Gazetteer

Tabel 3.1 Contoh Gazetter Nama Lokasi dan Tempat

idLoc	City	Street	Name
256	Surabaya	Jemursari	Hotel Santika Jemursari
712	Surabaya	Pucang Kerep	Hotel Pasah Asih
520	Surabaya	Nginden Intan Utara	Country Heritage Resort Hotel
769	Surabaya	Babatan Pratama	Perumahan Babatan Pratama

3.4.3 Praproses

Tahapan praproses dilakukan pada beberapa tahapan sistem. Jenis-jenis praproses yang digunakan pada sistem sebagai berikut: 1) *casefolding*: mengubah seluruh jenis huruf menjadi huruf kecil, 2) hapus url dan *mention*: menghapus substring dari link dan mention pada data tweet, 3) normalisasi: mengubah beberapa kata singkatan populer kebentuk panjangnya dan kata tidak baku menjadi bentuk kata baku, 4) menghapus tanda baca kecuali tanda tanya ‘ ? ‘, titik ‘ . ’, koma ‘ , ’, dan strip ‘- ‘. Digram alur tahap Praproses ditunjukkan pada Gambar 3.5.



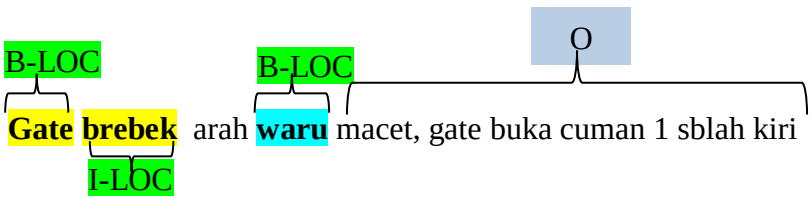
Gambar 3.5 Diagram Alur Tahapan Praproses

3.4.4 Ekstraksi informasi berbasis *NeuroNER*

Teknik ekstraksi informasi pada penelitian ini menggunakan *NeuroNER* untuk melakukan ekstraksi entitas. Jenis entitas yang digunakan pada penelitian ini dibagi menjadi enam label notasi entitas, antara lain: 1) LOC (Location), 2) GPE (Geograpical Political Entity), 3) HWYMSE (Highway Position), 4) BLD (Building), 5) NPL (Natural Place), dan 6) MSE (Measurement) seperti pada Tabel 3.2. Agar sistem dapat melakukan ekstraksi informasi dengan baik diperlukan data latih yang cukup besar dan pelabelan yang baik. Data latih yang digunakan sama seperti data latih untuk klasifikasi jenis tweet dan dilakukan pelabelan secara manual. Format pelabelan data yang digunakan merujuk pada [10][11] menggunakan format IOB atau BIO. Diberi label B (Begin) jika entitas merupakan kata pertama dari entitas, I (Inside) jika kata bukan merupakan kata pertama dari nama entitas, dan O (Outside) jika kata tidak termasuk dalam kategori entitas apapun. Contoh dari pelabelan data ditunjukkan pada Gambar 3.6 dan diagram alur NeuroNER ditunjukkan pada Gambar 3.7.

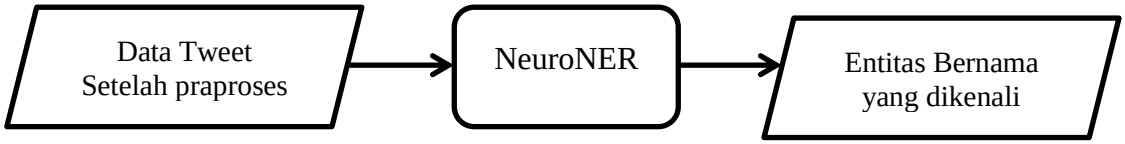
Tabel 3.2 Daftar Notasi Entitas Bernama

Label Entitas	**Contoh Entitas**
LOC (Location)	Wonokromo, Mulyorejo
GPE (Geograpical Political Entity)	Surabaya, Sidoarjo
HWYMSE(Highway Measurement)	Km 4, km 40.100
BLD (Building)	Swiss-Belinn, Tunjungan Plaza
NPL (Natural Place)	Sungai jagir, Gunung Bromo
OBJ	Truk patas as , Avanza
MSE (Measurement)	KM 90, 5.6 SR, 40 cm
TIME	19.30 , 19:30
DATE	1-Juli-2018 , 1/07/2018
O (Other)	Saya , kamu , dimana, macet

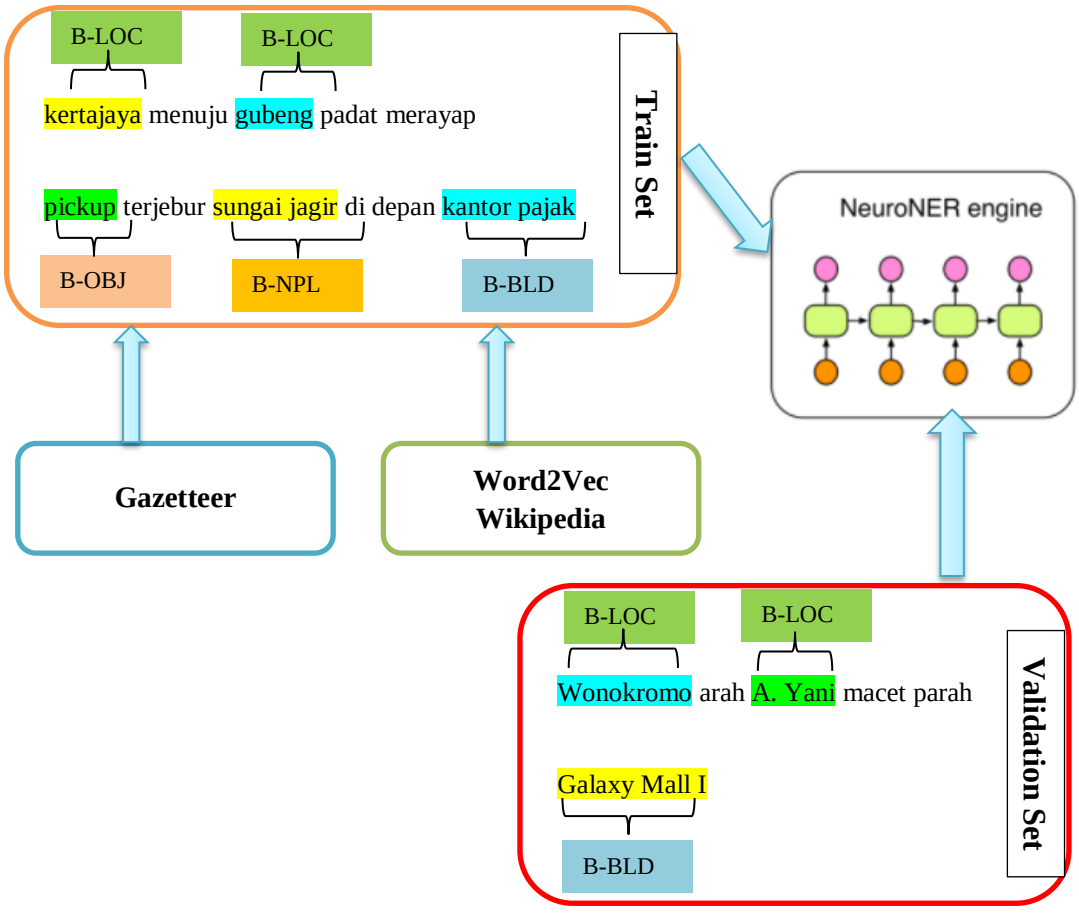


Gambar 3.6 Contoh pelabelan data latih NeuroNER

Selain penggunaan data latih berupa tweet yang telah dilabeli, tahap pelatihan *NeuroNER* juga ditambahkan data dari wikipedia yaitu seluruh artikel bahasa indonesia yang telah dibentuk menjadi word2vec dan penambahan data Gazetteer yang dapat meningkatkan performa pengenalan entitas bernama dari *NeuroNER*. Diagram Alur pelatihan *NeuroNER* ditunjukkan pada Gambar 3.8 sedangkan alur Testing ditunjukkan pada Gambar 3.7.



Gambar 3.7 Diagram Alur Ekstraksi Entitas Bernama dengan NeuroNER



Gambar 3.8 Diagram Alur Pelatihan NeuroNER

3.4.5 Klasifikasi Jenis Tweet Informasi Kejadian

Tahapan berikutnya setelah pengenalan entitas bernama yaitu deteksi jenis *tweet* Informasi menggunakan klasifier RCNN. Tahapan klasifikasi dilakukan setelah tahapan pengenalan entitas adalah karena merujuk pada definisi informasi kejadian penting yaitu harus mempunyai paling tidak satu entitas yang merepresentasikan lokasi. Maka dari itu tugas dari *NeuroNER* untuk mengenali entitas-entitas yang ada dalam tweet.

Pada dasarnya jenis tweet ada dua yaitu *tweet* informasi kejadian penting dan *tweet* non-informasi kejadian penting. Kemudian *tweet* informasi kejadian penting diklasifikasikan lagi menjadi tiga kelas yaitu lalu-lintas, kebakaran, dan bencana alam. Sehingga disini jenis tweet informasi kejadian diklasifikasikan menjadi empat kelas yaitu 1)Non-Informasi Kejadian Penting, 1) Lalu-Lintas, 2) Kebakaran, dan 3) Bencana-Alam. Tweet yang termasuk kedalam kelas informasi kejadian penting akan diteruskan ke tahapan proses selanjutnya. Sedangkan tweet yang masuk kedalam kelas non-informasi kejadian penting akan diabaikan atau tidak diteruskan ke tahapan proses selanjutnya. Contoh tweet untuk masing-masing kelas ditunjukkan pada Tabel 3.3.

Tabel 3.3 Contoh Tweet untuk masing-masing Jenis Informasi Kejadian

Jenis Informasi Kejadian	**Teks Tweet**
Lalu-Lintas	LOC arah LOC padat
Lalu-Lintas	Info awal kecelakaan OBJ terguling di LOC arah LOC
Kebakaran	TIME info awal kebakaran di lahan kosong jl LOC seberang showroom
Kebakaran	Kebakaran kapal di LOC
Bencana-Alam	sejumlah kendaraan terjebak banjir di LOC
Bencana-Alam	gempabumi tektonik mag MSE menguncang LOC
Non-Informasi Kejadian Penting	LOC arah LOC lanjar
Non-Informasi Kejadian Penting	desak investigasi kebakaran BLD dilanjutkan

3.4.6 Ekstraksi Informasi Berbasis *Relation Extraction* (RE)

Proses RE pada penelitian ini digunakan untuk melakukan ekstraksi relasi antar entitas bernama yang sudah dikenali oleh *NeuroNER*. Algoritma yang digunakan untuk RE adalah klasifier RCNN sehingga model RE disini disebut sebagai *NeuralRE*. Hasil ekstraksi relasi berupa label label relasi yang berguna untuk mengetahui hubungan entitas satu dengan entitas lainnya. Jumlah label relasi dan contohnya ditunjukkan pada Tabel 3.7.

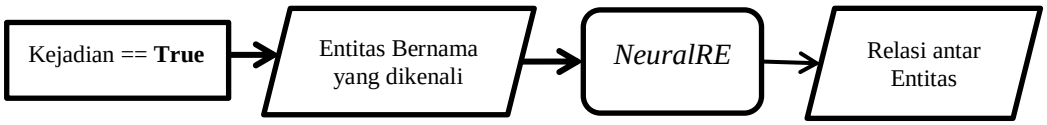
Terdapat batasan yang peneiti tetapkan pada proses RE yaitu hanya memproses entitas bernama yang berlabel (LOC, GPE, BLD, HWYMSE) dengan maksimal jumlah 4 entitas bernama dalam satu tweet, kecuali tweet yang sudah terdapat *substring* pemisah yang sudah ditetapkan. Tweet yang teridentifikasi dengan pemisah yang ditetapkan akan dilakukan pemisahan tweet sebanyak jumlah pemisah yang dikenali. Contoh dari tweet dengan jumlah lebih dari 4 entitas dan tweet dengan pemisah substring yang sudah ditetapkan ditunjukkan pada Tabel 3.3 dan Diagram alur RE ditunjukkan pada Gambar 3.9 sedangkan hasil ekstraksi ditunjukkan pada Tabel 3.4.

Tabel 3.4 Label Relasi dan Contohnya

Label Relasi	**Teks Tweet**
Highway-Position (e1,e2)	Macet di **tol sumo** arah mojokerto sampai **km 20** **Highway** **Position**
Highway-Position (e2,e1)	Macet di **km 20 tol sumo** arah mojokerto **Position** **Highway**
Street-Place (e1,e2)	**Ayani** macet di depan **royal plaza** **Street** **Place**
Street-Place (e2,e1)	Depan **royal** jl. **ayani** macet **Place** **Street**
StartingPoint-Destination (e1,e2)	**Ayani** arah **wonokromo** padat merayap **StartingPoint** **Destination**
StartingPoint-Destination (e2,e1)	**Ayani** dari arah **wonokromo** padat merayap **Destination** **StartingPoint**
Other	Kecelakaan di **embong malang** , korban dibawa ke **rsi** **Other**

Tabel 3.5 Contoh Tweet Multi-Entitas dan Tweet dengan Pemisah Khusus

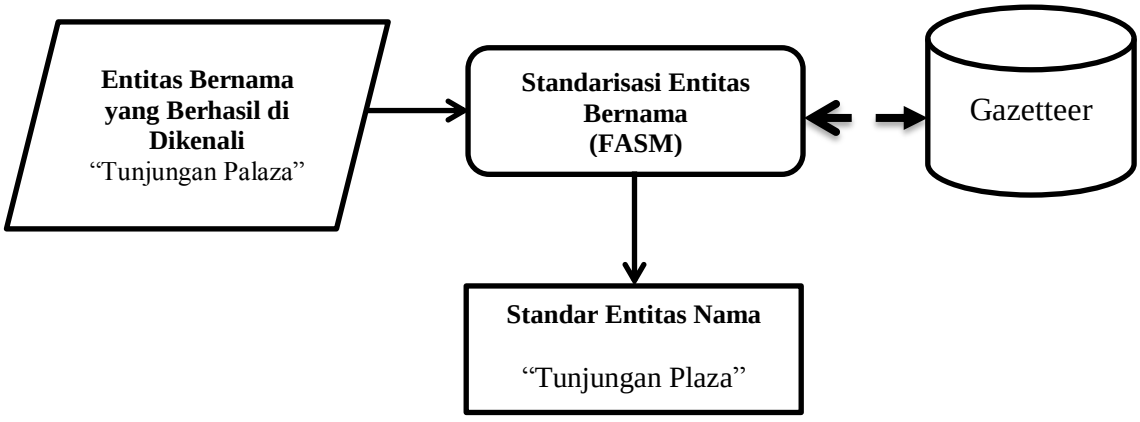
Tweet tanpa pemisah yang ditetapkan dengan jumlah entitas lebih dari empat		
"@Dwihariyatno2: **Gresik**: jl **mayjend sungkono** dr **nippon paint** sampai depan **rusunawa Surabaya** : jl **margomulyo**, jl **raya greges** sampai **romokalisari**"		
Tweet dengan pemisah yang ditetapkan	**Pemisah**	**Hasil**
16.45: Waspadai kepadatannya 1. Wonokusumo arah Kedung Mangu padat 2. Ngagel arah Jagir padat 3. Ketabang kali - Moestopo padat 4. Waru - Aloha 2 arah padat merambat (odp-pr)	1. 2. 3. 4.	Wonokusumo arah Kedung Mangu padat Ngagel arah Jagir padat Ketabang kali - Moestopo padat Waru - Aloha 2 arah padat merambat



Gambar 3.9 Diagram Alur Ekstraksi Relasi antar Entitas Bernama

3.4.7 Standarisasi Entitas Nama

Setelah entitas nama dikenali maka dilakukan validasi penulisan entitas disesuaikan dengan Gazetteer dan kamus kata entitas lainnya sehingga didapatkan entitas nama sesuai standar. Proses standarisasi nama dilakukan menggunakan teknik FASM yang ditunjukkan pada Gambar 3.10.



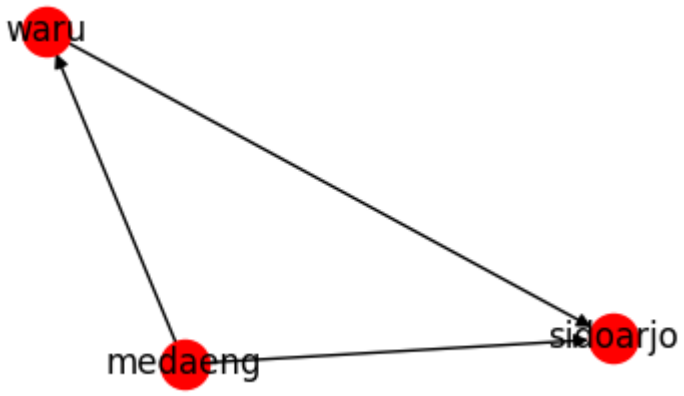
Gambar 3.10 Diagram Alur Standarisasi Entitas Nama

3.4.8 Konversi Hasil Ekstraksi Informasi ke Bentuk Graf

Merupakan proses mengubah hasil ekstraksi informasi khususnya entitas bernama yang tergolong sebagai unsur tempat kejadian. Entitas bernama yang termasuk dalam golongan unsur tempat kejadian antara lain : LOC, BLD, GPE, NPL, dan HWYMSE. Setiap kelima entitas bernama tersebut dikenali oleh NeuroNER maka entitas bernama tersebut akan dianggap sebagai node baru pada graf. Setiap node yang memiliki relasi startingPoint-Destination akan dihubungkan ke node lainnya sesuai hasil ekstraksi relasi oleh *NeuralRE*. Adapun masing-masing jenis relasi mempunyai perlakuan yang berbeda dalam pembentukan graf. Perlakuan masing-masing jenis relasi ditunjukkan pada Tabel 3.6.

Tabel 3.6 Perlakuan untuk masing-masing Relasi pada Graf

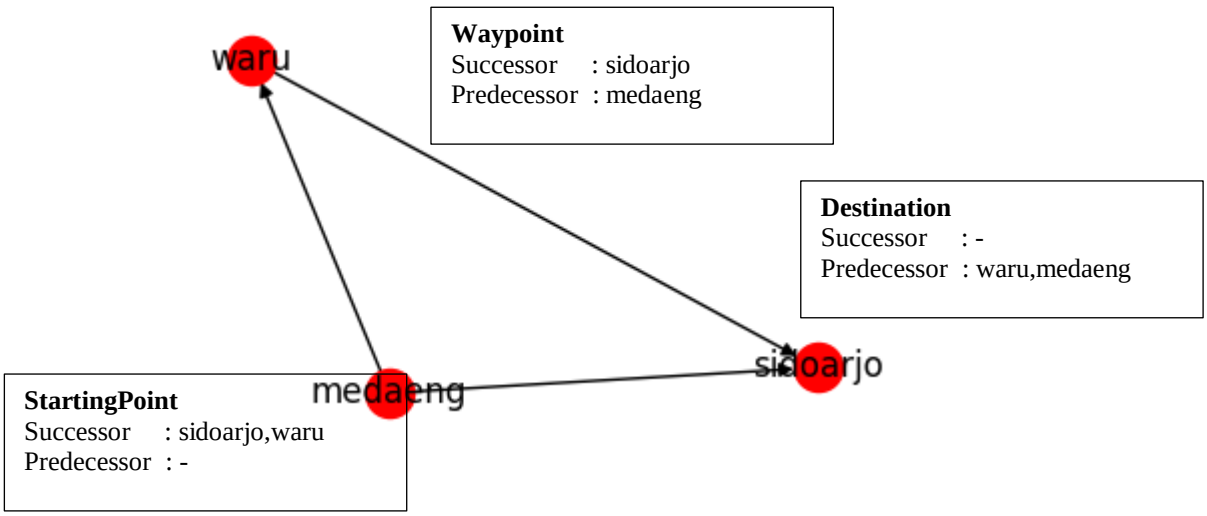
No	**Relasi**	**Perlakuan**
1	**Highway-Position**	Penggabungan Node antara Lokasi jalan Tol dengan Posisinya. Dan menghapus node posisi
2	**Street - Place**	Penggabungan Node antara Lokasi dengan Tempat. Dan menghapus node Tempat
3	**StartingPoint - Destination**	Node starttingPoint dihubungkan dengan Edge ke node Destination
4	**Other**	Tidak ada



Gambar 3.11 Contoh Konversi Hasil Ekstraksi Relasi ke Bentuk Graf

Node dan edge yang sudah didapat akan dikonversi ke dalam bentuk graf yaitu directed-graph. Contoh hasil ekstraksi relasi jika diubah ke bentuk graf seperti pada Gambar 3.11. Hal ini dilakukan untuk mendapatkan node sebagai titik kejadian jika node tersebut berupa node tunggal, dan node sebagai titik awal bermula dari kejadian jika node tersebut memiliki hubungan dengan node yang lain. Jenis node yang berhubugan dengan node lainnya dibagi menjadi tiga jenis node yaitu: 1) node sebagai s*tarting point*, 2) node sebagai *destination*, 3) node sebagai *waypoint*.

Untuk menentukan siapa node yang menjadi startingpoing, destination, dan waypoint, digunakan konsep graf dengan menganalogikan starting point merupakan root dari sebuah graf, destination sebagai node akhir dari cabang graf, dan waypoint adalah node yang menghubungkan antara root dengan node akhir. Sebuah node dapat dijadikan sebagai **root** atau **startingpoint** jika node tersebut tidak memiliki predecessor dan sebaliknya memiliki successor. Kemudian untuk menentukan node sebagai **destination** adalah memilih node yang tidak memiliki successor namun memiliki predecessor. Sedangkan node yang mempunyai predecessor dan successor adalah merupakan **waypoint**. Node yang sudah dikenali dan diketahui jenis nodenya maka akan disimpan ke dalam database sebagai inputan pada tahap visualisasi data. Contoh hasil proses penentuan startingpoint, destination, dan waypoints ditunjukkan pada Gambar 3.12.



Gambar 3.12 Contoh Konversi Hasil Ekstraksi Relasi ke Bentuk Graf setelah melalui Proses Penentuan *Startingpoint*, *Destination*, dan *Waypoint*

3.4.9 Filter Deteksi Kesamaan Informasi dan Pengelompokan Informasi Kejadian

Tahap terakhir dari sistem ini sebelum visualisasi data informasi kejadian adalah proses filter data informasi kejadian. Proses ini bertujuan untuk memilah informasi kejadian yang akan disimpan dalam database. Pengelompokan data kejadian, merupakan proses pegelompokan informasi kejadian berdasarkan kesamaan teks informasi, lokasi, dan jenis kejadian dengan rentang waktu kejadian tertentu.

Rentang waktu kejadian adalah selisih informasi awal dan informasi berikutnya yang merupakan satu informasi yang sama. Rentang waktu kejadian berupa satuan waktu yang ditetapkan peneliti sebesar 120 menit. Ada beberapa kriteria dalam proses *filtering* dan pengelompokan informasi kejadian yaitu antara lain sebagai berikut:

1. **Isi teks tweet sama persis** dengan tweet yang sudah masuk database, maka akan diabaikan oleh sistem.
2. **Isi teks tweet berbeda namun mengandung informasi yang sama dan atau informasi berbeda,** maka tweet akan dimasukan sebagai referensi tambahan dari sebuah informasi kejadian yang sama**.** Contoh filtering dan pengelompokan tweet untuk masing-masing kriteria ditunjukkan pada Tabel 3.7.

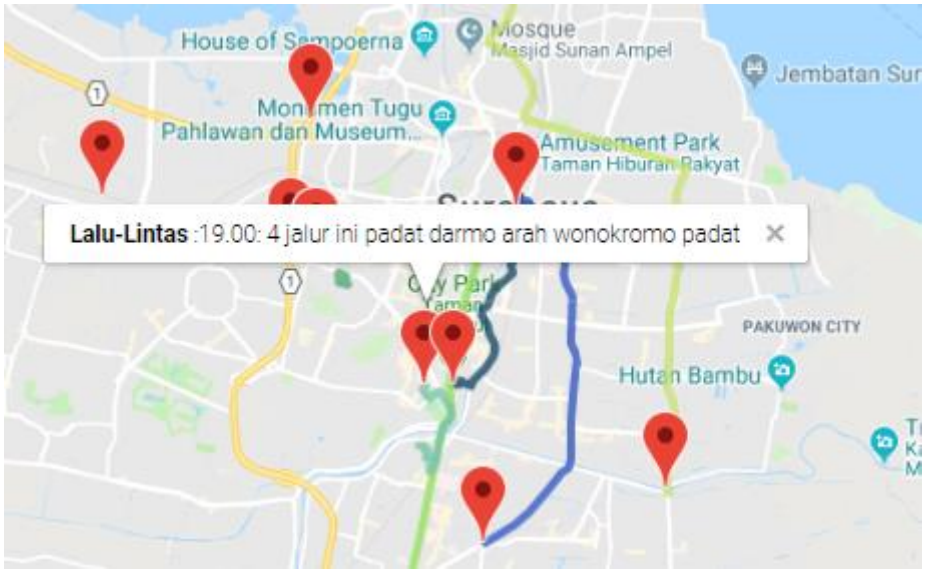
Tabel 3.7 Contoh Filter dan Pengelompokan Tweet

No. Kriteria	Teks Tweet Pertama	Teks Tweet Kedua
1	#InfoBMKG Gempa Mag 5.8, lokasi 161km sebelah tenggara Kabupaten Malang. Kedalaman 10km. Tidak berpotensi Tsunami. (odp-pr)	~~RT @e100ss:~~ #InfoBMKG Gempa Mag 5.8, lokasi 161km sebelah tenggara Kabupaten Malang. Kedalaman 10km. Tidak berpotensi Tsunami. (odp-pr)
2	Kebakaran kapal di **pelabuhan gresik**	Info awal terjadi kebakaran kapal di **Pelabuhan Gresik** sekitar pukul 10.30. Sudah ada petugas, belum diketahui penyebabnya

3.4.10 Visualisasi Data

Tahap terakhir dari sistem ini adalah visualisasi data. Tujuan dari visualisasi data adalah untuk memudahkan masyarakat dalam membaca informasi kejadian yang ada. Sistem akan menampilkan semua informasi yang masuk kedalam database yang waktu kejadiannya terjadi 90 menit sebelum waktu akses. Peneliti menggunakan *tool* bantuan Google Maps Api V3 untuk memudahkan visualisasi data. Data yang telah disimpan di database pada tahap sebelumnya akan dijadikan input visualisasi data.

Untuk masing-masing lokasi kejadian akan diwakili dengan sebuah marker jika tidak terdapat titik destinasi. Jika lokasi kejadian mempunyai informasi titk awal dan titik tujuan, maka titik awal akan diwakili oleh marker dan dihubungkan ke titik tujuan dengan garis (*direction*). Warna marker dibedakan untuk masing-masing jenis informasi kejadian. Marker warna merah untuk jenis informasi kejadian lalu-lintas, marker kuning, untuk jenis kejadian kebakaran, dan marker warna hijau untuk jenis informasi kejadian bencana-alam. Contoh hasil dari visualisasi data informasi kejadian ditunjukkan pada Gambar 3.13.



Gambar 3.13 Contoh Hasil Visualisasi Data Informasi Kejadian

3.5 Ujicoba dan Analisa Hasil

Setelah tahapan desain sistem dibuat maka selanjutnya dilakukan ujicoba sistem seperti pada Gambar 3.2 untuk mengetahui kemampuan gabungan metode yang diusulkan untuk deteksi kejadian pada data stream Twitter yang meliputi performa pengenalan entitas bernama, klasifikasi jenis tweet informasi, ekstraksi relasi, dan standarisasi nama Gazetter. Metode pengujian yang digunakan untuk menguji performa sistem secara keseluruhan adalah perhitungan sensitivity. Sedangkan metode uji yang digunakan untuk menguji performa pengenalan entitas bernama, klasifikasi informasi kejadian, dan relasi ekstraksi adalah perhitungan *precision*, *recall*, dan *F-measure*.

3.4.1 Sensitivity

Sensitivity digunakan untuk menghitung performa keberhasilan sistem secara keseluruhan dari tahap awal sampai tahap terakhir. Persamaan yang digunakan seperti ditunjukkan pada Persamaan 3.1. Dimana, TP = *true positif*, FN = f*alse negatif*, dan TP adalah jumlah keberhasilan sistem dalam mengenali entitas, klasifikasi kejadian, dan ekstraksi relasi. Sedangkan, FN adalah jumlah ketidak berhasilan sistem dalam mengenali entitas, klasifikasi kejadian, dan atau ekstraksi relasi.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \; x \; 100\% \tag{3.1}$$

3.4.2 Precision, Recall, dan F-Measure

3.4.2.1 Pengenalan Entittas Bernama

Pada proses NeuroNER, precision digunakan untuk mengukur kemampuan sistem pengenalan entitas bernama dalam mengenali jenis entitas dengan menghitung perbandingan jumlah entitas yang secara benar dikenali dengan jumlah seluruh entitas yang dikenali. Rumus precision yang digunakan seperti pada Persamaan 3.2. TP = *true positive*, FP = *false positive*, dimana TP adalah jumlah entitas yang berhasil dikenali dengan benar. Sedangkan, FP adalah jumlah entitas yang dikenali dan terkenali dengan salah dalam entitas tertentu.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \; x \; 100 \tag{3.2}$$

Recall untuk mengukur kemampuan sistem untuk mengenali entitas yang relevan dalam jenis entitasnya. Rumus recall yang digunakan seperti pada Persamaan 3.3, dimana, FN = *false negatif*, yaitu jumlah kesalahan pengenalan entitas yang dikenali dalam jenis entitasnya.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \; x \; 100\% \tag{3.3}$$

F-Measure **(F)** adalah *harmonic mean* dari *precision* dan *recall.* Rumus f-measure seperti pada Persamaan 3.4.

$$F = 2 \; x \; \frac{precision \; x \; recall}{precision + recall} \; x \; 100\% \tag{3.4}$$

3.4.2.2 Klasifikasi Jenis Kejadian Penting

Pada proses klasifikasi, precision digunakan untuk mengukur kemampuan sistem dalam melakukan klasifikasi jenis kejadian dengan menghitung perbandingan jumlah kelas yang secara benar dikenali dengan jumlah seluruh kelas yang dikenali. Rumus precision yang digunakan seperti pada Persamaan 3.5, dimana, TP = *true positive*, FP = *false positive*, dan TP adalah jumlah kejadian yang berhasil diklasifikasikan dengan benar. Sedangkan, FP adalah jumlah kejadian yang dikenali dan diklasifikasikan dengan salah.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \; x \; 100 \tag{3.5}$$

Recall untuk mengukur kemampuan sistem untuk mengklasifikasikan kejadian secara benar dalam kelasnya. Rumus recall yang digunakan seperti pada Persamaan 3.6, dimana, FN = *false negatif*, yaitu jumlah kesalahan klasifikasi kejadian dalam kelas entitas tertentu.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \; x \; 100\% \tag{3.6}$$

***F-Measure* (F)** adalah *harmonic mean* dari *precision* dan *recall*. Rumus f-measure seperti pada Persamaan 3.7.

$$F = 2 \; x \; \frac{precision \; x \; recall}{precision + recall} \; x \; 100\% \tag{3.7}$$

3.4.2.3 Ekstraksi Relasi antar Entitas

Pada proses ekstraksi relasi, precision digunakan untuk mengukur kemampuan sistem dalam melakukan ekstraksi jenis relasi antar entitas dengan menghitung perbandingan jumlah jenis relasi yang secara benar dikenali dengan jumlah seluruh kelas relasi yang dikenali. Rumus precision yang digunakan seperti pada Persamaan 3.8, dimana, TP = *true positive*, FP = *false positive*, dan TP adalah jumlah relasi yang berhasil di-ekstrak dengan benar. Sedangkan, FP adalah jumlah jenis relasi yang dikenali dan diklasifikasikan dengan salah.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \; x \; 100 \tag{3.8}$$

Recall untuk mengukur kemampuan sistem untuk mengekstrak relasi secara benar dalam kelasnya. Rumus recall yang digunakan seperti pada Persamaan 3.9, dimana, FN = *false negatif*, yaitu jumlah kesalahan ekstraksi relasi dalam kelas tertentu.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \; x \; 100\% \tag{3.9}$$

***F-Measure* (F)** adalah *harmonic mean* dari *precision* dan *recall*. Rumus f-measure seperti pada Persamaan 3.10.

$$F = 2 \; x \; \frac{precision \; x \; recall}{precision + recall} \; x \; 100\% \tag{3.10}$$

BAB 4
Uji Coba dan Analisa Hasil

Pada bab ini akan dijabarkan hasil dari implementasi dari masing-masing tahapan yang telah dijelaskan pada Bab 3. Berikutnya dijabarkan hasil evaluasi dari ujicoba dengan metode evaluasi perhitungan *sensitivity*, *precision*, *recall*, dan *f-measure*.

4.1 Lingkungan Implementasi

Metode yang digunakan dalam penelitian ini diimplementasikan dengan dukungan hardware sebuah *personal computer* (PC) dengan spesifikasi prosesor intel core i5-2.3 GHz, RAM 8 GB, dan didukung sistem operasi windows 10.1 dengan menggunakan bahasa pemrograman Python 3.6. Database yang digunakan untuk meyimpan data twitter dan hasil ujicoba disimpan di database menggunakan MySQL. Untuk library utama yang digunakan dalam penelitian ini adalah adalah tensorflow yang membantu proses ekstraksi informasi menggunakan algoritma *deep-learning*.

4.2 Deskripsi Data

Data penelitian dibagi menjadi dua yaitu data latih dan data uji coba. Semua data yang digunakan berupa tweet yang diambil dari Twitter secara bebas dan gratis. Pengambilan data dibagi menjadi dua cara yaitu *crawling* dan *streaming*. Crawling untuk mendapatkan data yang digunakan sebagai data latih, dan streaming untuk mendapatkan data yang digunakan untuk uji coba sistem.

4.2.1 Data Latih

Data diambil dengan cara *crawling* menggunakan API Twitter berdasarkan dari spesifik user_id pengguna dan berdasarkan kata kunci. Untuk pengambilan data pada spesifik *timeline* dari pengguna yang diamati adalah akun Twitter dari Radio Suara Surabaya (@e100ss), BMKG (@infoBMKG), dan Dishub Surabaya (@sits_dishubsby). Sedangkan pengambilan tweet berdasarkan kata kunci digunakan beberapa kata yaitu ‘kecelakaan’, ‘kebakaran’, ‘bencana’,

‘gempa’, dan ‘macet’. Data diambil pada kurun waktu empat bulan sekitar bulan Maret – Juni.

Data tweet yang didapatkan terdiri dari dua unsur yaitu; 1) teks tweet, dan 2) waktu tweet. Contoh data seperti yang ditunjukkan pada Tabel 4.1. Kemudian data disimpan pada file *plain text* sebelum dilakukan praproses.

Tabel 4.1 Tabel Contoh Data Mentah *Tweet* atau Postingan dari Twitter

Waktu Tweet	**Teks Tweet**
2018-03-10 14:21:17	RT @danielht2009: Banjir di Jln Dr. Sutomo. @e100ss https://t.co/FugqxQru5S
2018-03-10 11:50:25	RT @stvchenx: @e100ss wiyung arah unesa macet total https://t.co/koC5n1l8i8
2018-03-10 13:17:42	RT @aisiyahdisri: Posisi lalu lintas di depan Royal Plaza, padat merayap @e100ss https://t.co/sKkoFl2POq
2018-03-08 05:34:20	terjadi kecelakaan dijalur mantan, seorang jomblo terluka parah dibagian hati
2018-06-17 09:28:20	Info update ttg kebakaran di timur Pasar Gadang - Malang. Sudah terkondisikan. Kebakaran di bawah jembatan Gadang. Dlm pendinginan saat ini @RadioElshinta @PuspitaFM @Senaputra_FM @infomalangraya @infomalang @LalinNews @e100ss

Proses-proses yang memerlukan data latih antara lain: 1) *NeuroNER*, 2) Klasifikasi jenis informasi kejadian, dan 3) *NeuralRE*.

4.2.1.1 Data Latih NeuroNER

Data tweet yang sudah dilakukan praproses dilakukan pelabelan untuk setiap token sesuai dengan nama entitas bernama dari masing-masing token. Jumlah data *tweet* yang digunakan sebagai data latih *NeuoNER* sebanyak 1.114 *tweet* atau sebanyak 16.845 token. Rincian jumlah token untuk masing-masing jenis entitas bernama ditunjukkan pada Tabel 4.2 dan Contoh pelabelan untuk data latih *NeuroNER* ditunjukkan pada tabel 4.3.

Tabel 4.2 Jumlah Data Latih *NeuroNER*

Nama Entitas	Format BIO	Jumlah
LOC	B-LOC	988
	I-LOC	911
GPE	B-GPE	333
	I-GPE	55
BLD	B-BLD	247
	I-BLD	287
HWYMSE	B-HWYMSE	37
	I-HWYMSE	38
NPL	B- NPL	18
	I- NPL	22
OBJ	B- OBJ	518
	I- OBJ	672
MSE	B- MSE	230
	I- MSE	260
TIME	B-TIME	193
	I-TIME	41
DATE	B-DATE	79
	I-DATE	58
Other	O	11.851
Total		16.845

Tabel 4.3 Pelabelan Data Latih *NeuroNER*

Teks Tweet	Token	Label (BIO)
Basuki rahmad arah tunjungan plaza macet	basuki rahmad arah tunjungan plaza macet	B-LOC I-LOC O B-BLD I-BLD O
tol gempan arah arteri porong padat	tol gempan arah arteri porong padat	B-LOC I-LOC O B-LOC I-LOC O

4.2.1.2 Data Latih Klasifikasi Jenis Kejadian Penting

Data *tweet* hasil *crawling* dilakukan praproses antara lain; 1) remove URL dan mention, 2) casefolding, 3) merubah singkatan dan kata tidak baku kebentuk umum, 4) menghapus huruf berulang pada kata, 3) menghapus semuatanda baca kecuali tanda strip ‘ – ‘ yang memiliki makna yang sama dengan kata ‘arah’ dan ‘menuju’. Selain itu model dari data latih untuk klasifikasi adalah teks tweet setelah melalui proses pengenalan entitas bernama oleh *NeuroNER*. Seperti yang sudah dijelaskan di Bab 3, jumlah kelas atau jenis kejadian dibagi menjadi empat yaitu; 1) Non-Informasi Kejadian Penting, 2) Lalu-Lintas, 2) Kebakaran, 3) Bencana-Alam. Jumlah data latih yang digunakan sebanyak 776 *tweet* dengan rincian setiap jenisnya seperti pada Tabel 4.5 dan contoh dari data latih untuk pembentukan model data latih klasifikasi informasi kejadian penting ditunjukkan pada Tabel 4.4.

Tabel 4.4 Tweet Data Latih dan Label Jenis Informasi Kejadian

Teks Tweet	Data Latih Klasifikasi Jenis Kejadian	Label Kelas
wiyung arah unesa macet total	LOC arah BLD macet total	Lalu-Lintas
tol gempan arah arteri porong padat	LOC arah LOC padat	Lalu-Lintas
kebakaran pabrik di daerah geluran taman	kebakaran pabrik di daerah LOC	Kebakaran
goyangan gempa bumi di malang	goyangan gempa bumi di GPE	Bencana-Alam
terjadi kecelakaan dijalur mantan, seorang jomblo terluka parah dibagian hati	terjadi kecelakaan dijalur mantan, seorang jomblo terluka parah dibagian hati	Non-Informasi Kejadian Penting
ahmad yani arah wonokromo lancar jaya	LOC arah LOC lancar jaya	Non-Informasi Kejadian Penting

Tabel 4.5 Jumlah Data Latih Klasifikasi Jenis Informasi Kejadian

Jenis Informasi Kejadian	Jumlah
Informasi Non-Kejadian Penting	428
Lalu-Lintas	150
Kebakaran	122
Bencana-Alam	114
Total	814

4.2.1.3 Data Latih *Neural Relation Extraction*

Untuk data latih *NeuralRE* hampir sama dengan data latih untuk klasifikasi jenis informasi kejadian penting yang berupa data keluaran dari *NeuroNER* karena memang fungsi dari *NeuralRE* adalah untuk mendapatkan relasi antar entitas yang telah dikenali oleh *NeuroNER*. Namun, Sedikit berbeda dengan data latih klasifikasi informasi kejadian, pada data latih *NeuralRE* dilakukan penambahan *substring* untuk penanda antar entitas. Karena pada dasarnya relasi adalah hubungan antara dua entitas maka dari itu ditambahkan penanda sebagai fitur penting untuk *relation classification* yaitu berupa *substring* <e1> sebagai penanda entitas kesatu dan <e2> sebagai penanda entitas kedua.

Jumlah data latih yang digunakan untuk data latih *NeuralRE* dan rincian jumlah data untuk setiap jenis relasinya ditunjukkan pada Tabel 4.6, sedangkan contoh dari data latih NeuralRE ditunjukkan pada Tabel 4.7.

Tabel 4.6 Jumlah Data Latih *NeuralRE*

Kode Relasi	Jenis Relasi	Jumlah
1	Highway-Position(e1,e2)	16
2	Highway-Position(e2,e1)	8
3	Street-Place(e1,e2)	35
4	Street-Place(e2,e1)	7
5	StartingPoint-Destination(e1,e2)	160
6	StartingPoint-Destination(e2,e1)	71
0	Other	24
Total		321

Tabel 4.7 Contoh Data Latih *NeuralRE* dan Label Relasi

No	Data Latih NeuralRE	Label Relasi [kode relasi]
1	**<e1>LOC<e1>** arah **<e2>BLD<e2>** macet total	StartingPoint-Destination(e1,e2)
2	**<e1>LOC<e1>** arah **<e2>LOC<e2>** padat	StartingPoint-Destination(e1,e2)
3	kebakaran **<e1>BLD<e1>** di daerah **<e2>LOC<e2>**	Street-Place(e2,e1)
4	TIME : info awal : OBJ di **<e1>HWYMSE<e1>** di **<e2>LOC<e2>** - GPE . lalu lintas padat sejak **loc**	Highway-Position(e2,e1)
5	**<e1>LOC<e1>** - **LOC** agak tersendat di **<e2>HWYMSE<e2>** ada OBJ berhenti di lajur kanan, bagian depan penyok	Highway-Position(e1,e2)
6	**LOC** - **<e1>LOC<e1>** agak tersendat di **<e2>HWYMSE<e2>** ada OBJ berhenti di lajur kanan, bagian depan penyok	Other

Pada Tabel 4.7 baris ke-4, ke-5, dan ke-6 adalah contoh sebuah tweet yang memiliki lebih dari dua entitas dan lebih dari satu relasi. Untuk jenis tweet yang memiliki lebih dari dua entitas penanda lokasi seperti LOC, GPE, BLD, NPL, dan HWYMSE maka untuk data latih akan dibuatkan label relasi sebanyak kombinasi dua entitas dari semua entitas yang ada. Contoh pelabelan data untuk tweet dengan lebig dari 2 entitas bernama ditunjukkan pada Tabel 4.8.

Tabel 4.8 Contoh Pelabelan Tweet dengan Entitas Tempat Lebih dari Dua

NO	Data Latih NeuralRE	Label Relasi
1	**<e1>LOC<e1>** - **<e2>LOC<e2>** agak tersendat di **HWYMSE** ada OBJ berhenti di lajur kanan, bagian depan penyok	StartingPoint-Destination(e1,e2)
2	**<e1>LOC<e1>** - LOC agak tersendat di **<e2>HWYMSE<e2>** ada OBJ berhenti di lajur kanan, bagian depan penyok	Highway-Position(e1,e2)
3	LOC - **<e1>LOC<e1>** agak tersendat di **<e2>HWYMSE<e2>** ada OBJ berhenti di lajur kanan, bagian depan penyok	Other

4.2.2 Data Uji Coba

Data yang digunakan untuk uji coba sistem adalah berupa data stream yang diambil secara *streaming* menggunakan *stream* API Twitter secara gratis dan bebas. Akun Twitter yang diamati dan diambil data tweetnya adalah akun Radio Suara Surabaya dan Dishub Surabaya. Waktu pengambilan data bersamaan dengan waktu uji coba sistem yaitu pada tanggal 3 - 20 Juli 2018.

4.3 Uji Coba dan Hasil

Pengujian sistem menggunakan data tweet dengan cara mendapatkan data yang berbeda dengan tahapan pengambilan data untuk pelatihan sistem. Data diambil dengan cara *stream* seperti yang sudah dipaparkan pada subbab Data Uji Coba dan Gambar 3.4 yang merupakan diagram alur keseluruhan dari sistem. Namun, karena pada masa uji coba tidak semua data dapat menguji keseluruhan jenis kelas dari jenis entitas, jenis informasi kejadian dan jenis relasi, maka dilakukan re-posting atau tweet ulang data lama oleh akun peneliti.

Hasil pengujian dibagi menjadi menjadi 4 empat) yaitu; 1) ***precision, recall, dan f-measure***, untuk menghitung performa hasil pengenalan entitas bernama, 2) ***precision, recall, dan f-measure***, untuk menghitung performa hasil klasifikasi jenis informasi kejadian, 3**)** ***precision, recall, dan f-measure***, untuk menghitung performa hasil ekstraksi relasi antar entitas, dan 4) **Sensitivity**, menghitung persentase keberhasilan sistem dalam melakukan ketiga poin pengujian tersebut.

4.3.1 Skenario Uji Coba

Berdasar pada penjelasan di atas, tahap uji coba dilakukan pada tanggal 3 – 20 Juli 2018 pukul 08.00 - 23.00 wib. Hasil ujicoba pada setiap prosesnya dijelaskan sebagai berikut:

4.3.1.1 Praproses

Pada tahap praproses, data mentah berupa data stream dari Twitter dilakukan beberapa jenis praproses antara lain:

1) Menghapus semua tanda baca selain koma (,), titik(.), strip(-), garis miring(/), dan tanda tanya(?)

2) mengubah singkatan populer nama tempat di Twitter menjadi bentuk panjangnya, seperti ‘TP’ menjadi ‘Tunjungan Plaza’ dan ‘a.yani’ menjadi ‘Ahmad Yani’, dan
3) mengubah singkatan penanda nama tempat seperti ‘per3an’ menjadi pertigaan, ‘simpang 4’ menjadi ‘perempatan’.
4) menghapus ganti baris pada tweet.
5) menghapus kata ‘RT’ diawal teks tweet,
6) menghapus substring URL dan mention,
7) memberi jarak satu spasi antara teks alfabet dengan tanda baca.
8) Pemenggalan teks tweet menjadi beberapa tweet informasi bagi tweet yang terdapat tanda pemisah khusus yang sudah didefinisikan yaitu ‘1. ‘, ‘2. ‘, ‘3. ‘, ‘4. ‘, ‘5. ‘, dan ‘6. ‘.
9) Casefolding, merubah seluruh huruf dalam tweet menjadi huruf kecil

Contoh hasil dari setiap jenis praproses ditunjukkan pada Tabel 4.9.

Tabel 4.9 Contoh Hasil dari masing-masing Jenis Praproses

Urutan Praproses	**Input**	**Output**
1	Kapal Layar Motor **(KLM)** Sinar Timur terbakar di Pelabuhan Gresik, sekitar pukul 10.30 WIB, Senin **(16/7/2018**).	Kapal Layar Motor **KLM** Sinar Timur terbakar di Pelabuhan Gresik, sekitar pukul 10.30 WIB, Senin **16/7/2018**.
2	**A.Yani** mulai depan Kejaksaan arah Wonokromo padat	**Ahmad yani** mulai depan Kejaksaan arah Wonokromo padat
3	TL Per3an Sukomanunggal-Tol Banyu Urip-Simo mati @e100ss	TL Pertigaan Sukomanunggal-Tol Banyu Urip-Simo mati @e100ss
4	15.48: 3 jalur ini padat 1. Tol Sidoarjo arah Porong padat 2. Exit Tol Porong padat 3. Kalianak arah Margomulyo padat (odp-pr)	15.48: 3 jalur ini padat 1. Tol Sidoarjo arah Porong padat 2. Exit Tol Porong padat 3. Kalianak arah Margomulyo padat (odp-pr)
5	**RT** @S4R4SHadi: @e100ss raya grati sebelum per3an grati macet info ada kendaraan terguling	@S4R4SHadi: @e100ss raya grati sebelum pertigaan grati macet info ada kendaraan terguling
6	**@S4R4SHadi**: **@e100ss** raya grati sebelum pertigaan grati macet info ada kendaraan terguling **https://t.co/vi7Z9AMfog**	raya grati sebelum pertigaan grati macet info ada kendaraan terguling

Urutan Praproses	Input	Output
7	Exit Tol **Waru-Juanda MACET,** juga imbas **volume;**	Exit Tol **Waru - Juanda MACET ,** juga imbas **volume ;**
8	**16.45: Waspadai kepadatannya 1.** Wonokusumo arah Kedung Mangu padat **2.** Ngagel arah Jagir padat **3.** Ketabang kali - Moestopo padat **4.** Waru - Aloha 2 arah padat merambat	• Wonokusumo arah Kedung Mangu padat • Ngagel arah Jagir padat • Ketabang kali - Moestopo padat • Waru - Aloha 2 arah padat merambat
9	16.45: **Waspadai** kepadatannya **Wonokusumo** arah **Kedung Mangu** padat	16.45: **waspadai** kepadatannya **wonokusumo** arah **kedung mangu** padat

4.3.1.2 Pengenalan Entitas Bernama dengan *NeuroNER*

Dari hasil praproses dilanjutkan ke proses pengenalan entitas bernama menggunakan metode NeuroNER. Pengenalan entitas bernama dilakukan pada tahap awal sesudah praproses untuk mendapatkan syarat utama dari sebuah informasi kejadian penting yaitu mempunyai entitas lokasi kejadian.

Serangkaian praproses diatur sedemikian rupa untuk meningkatkan performa *NeuroNER* dalam melakukan pengenalan entitas bernama. Praproses yang berperan penting dalam ekstraksi entitas bernama adalah praproses ketujuh, yaitu memberi jarak satu spasi antara teks alfabet dengan tanda baca. Hal ini dilakukan karena data tweet yang masuk ke *NeuroNER* akan dilakukan tokenisasi dengan pemisah berupa spasi. Sehingga jika praprose ketujuh tidak dilakukan maka setiap kata yang dipisahkan hanya dengan tanda baca akan dihitung satu token seperti ditunjukkan pada Tabel 4.11. Terbacanya dua atau lebih kata menjadi satu token berakibat pada kesalahan pengenalan jenis entitas bernama oleh sistem.

Tabel 4.10 Hasil Proses Pengenalan Entitas Bernama dengan NeuroNER

No	Input	Output
1	**16.45:** Waspadai kepadatannya **Wonokusumo** arah **Kedung Mangu** padat	**TIME** Waspadai kepadatannya **LOC** arah **LOC** padat
2	depan **galaxy mall** masih macet krn pembangunan jembatan penyeberangan .	depan **BLD** masih macet krn pembangunan jembatan penyeberangan .
3	**km 12 perak** arah **waru** , pas naik **jembatan gunung sari** , **truk ngeban** jalur paling kiri , jadi padat	**HWYMSE LOC** arah **LOC** , pas naik **LOC** , **OBJ** jalur paling kiri , jadi padat
4	guyuran hujan abu vulkanik **gunung agung** terasa hingga **jember**	guyuran hujan abu vulkanik **NPL** terasa hingga **GPE**
5	**17.23** masih proses evakuasi . **truk as patah** jalur **surabaya probolinggo** . . antrian probolinggo arah **surabaya** sekitar **2 km**	**TIME** masih proses evakuasi . **OBJ** jalur **GPE GPE** . . antrian **probolinggo** arah **surabaya** sekitar **MSE**

Tabel 4.11 Kesalahan Pengenalan Jenis Entitas Bernama tanpa Praproses Ketujuh

Input	Token	Jenis Entitas
18.00 : malang-surabaya dan sebaliknya masih macet total.	18.00 : malang-surabaya dan sebaliknya masih macet total.	B-TIME O B-LOC O O O O
aloha....arah surabaya padat merayap	**aloha....**arah surabaya padat merayap	O B-GPE O O

Beberapa contoh hasil uji coba ditunjukkan pada Tabel 4.10 dan hasil percobaan *NeuroNER* dicatat dalam tabel confussion matrix seperti pada Tabel 4.12. Berdasarkan hasil pencatatan pada confussion matrix dapat dihitung nilai

precision, recall, dan *f-measure* untuk metode *NeuroNER*. Hasil perhitungan *precision, recall,* dan *f-measure* ditunjukkan pada Tabel 4.13.

Tabel 4.12 Tabel Confussion Matrix untuk Hasil *NeuroNER*

Label	LOC	GPE	BLD	HWYMSE	NPL	TIME	DATE	MSE	OBJ	Other
LOC	226							1		6
GPE		72							1	
BLD	1		35							
HWYMSE	1			20						
NPL					6					
TIME						65		1		
DATE						1	27	1		
MSE						1	2	34		1
OBJ			1					1	67	
Other	8					1	1	2	2	2000

Tabel 4.13 Precision, Recall, Dan F-Measure untuk Hasil *NeuroNER*

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**

4.3.1.3 Klasifikasi Jenis Kejadian dengan RCNN

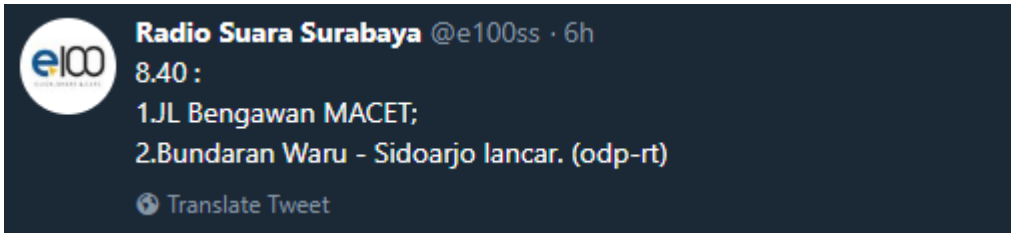
Hasil keluaran dari tahap pengenalan entitas bernama akan menjadi data masukan untuk tahap klasifkasi jenis kejadian, jika tweet terdeteksi mempunyai **minimal satu entitas bernama** seperti LOC, GPE, BLD, dan atau NPL. Contoh hasil proses klasifikasi jenis informasi kejadian ditunjukkan pada Tabel 4.14

Tabel 4.14 Hasil Proses Klasifikasi Jenis Informasi Kejadian dengan RCNN

No	Input	Jenis Kejadian
1	**TIME** Waspadai kepadatannya **LOC** arah **LOC** padat	Lalu-Lintas
2	depan **BLD** masih macet krn pembangunan jembatan penyeberangan .	Lalu-Lintas
3	dishub **TIME** wib arus lalu lintas simpang **LOC - LOC** terpantau lancar .	Non-Informasi Kejadian
4	untuk bapak2 kalo mau belok tapi lampu sen motornya mati coba kasi tanda pake lambaian tangan seperti pas tes sim	Non-Informasi Kejadian
5	kebakaran **OBJ** di **LOC - LOC HWYMSE**	Kebakaran
6	**TIME** info awal kebakaran di **LOC** . agus mengirim foto dan info yang terbakar adalah kapal barang dan apinya besar . asapnya hitam membubung tinggi . banyak warga pekerja pelabuhan di lokasi . info sudah diteruskan ke petugas .	Kebakaran
7	gempa mag **MSE DATE TIME** wib lok 0.24 lu 122.09 bt **MSE** baratdaya **LOC - GPE** , kedlmn **MSE** bmkg (gempa mag 5.1 sr 10- jul - 18 22 33 21 wib lok 0.24 lu 122.09 bt 54 km baratdaya boalemo - gorontalo , kedlmn 186 km bmkg)	Bencana-Alam

Praproses yang berperan penting untuk memaksimalkan hasil klasifikasi jenis informasi kejadian adalah praproses kedelapan. Adanya Pola penulisan tweet dengan jenis informasi kejadian yang berbeda lokasi namun dengan waktu kejadian yang sama dalam satu tweet menjadi alasan penggunaan praproses ketujuh. Contoh

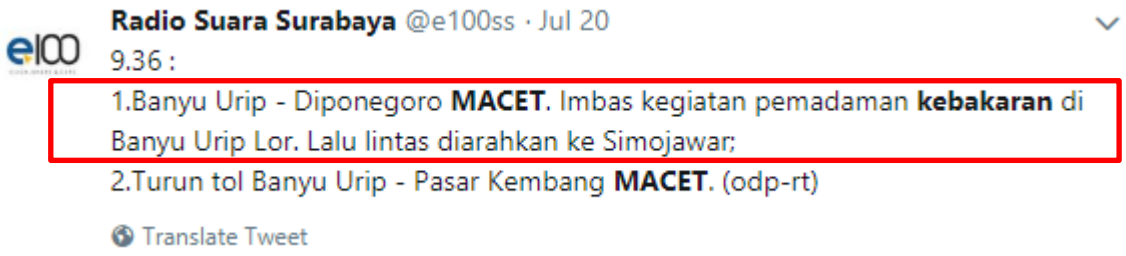
penulisan informasi berbeda jenis informasi kejadian dalam satu tweet seperti ditunjukkan pada Gambar 4.1.



Gambar 4.1 Satu Tweet Berisi Dua Jenis Informasi Berbeda dengan pemisah angka

Jika pemisahan tweet berdasarkan pemisah khususnya yaitu angka dan titik (1. dan 2.) tidak dilakukan maka akan terjadi kesalahan dalam klasifikasi. Maka dari itu perlu dilakukan pemisahan satu tweet menjadi beberapa tweet sesuai penanda pemisah khusus yang dikenali. Untuk contoh Gambar 4.1, tweet akan dipisah menjadi dua tweet yaitu; 1) Jl. Bengawan Macet; dan 2) Bundaran Waru – Sidoarjo lancar. (odp-rt). Sehingga pada **tweet pertama** dapat diklasifikasikan sebagai kelas kejadian Lalu-Lintas dan **tweet kedua** sebagai kelas Non-Informasi Kejadian Penting.

Pada kasus satu tweet memiliki lebih dari saru jenis informasi kejadian seperti pada Gambar 4.2, yang menunjukkan bahwa sedang terjadi kemacetan (Lalu-Lintas) dan kebakaran di Banyu Urip, sistem akan tetap mengklasifikasikan *tweet* pada salah satu jenis kejadian. Hal tersebut karena tipe klasifier yang digunakan hanya untuk klasifikasi single-label sehingga tidak mampu melakukan klasifikasi multi-label.



Gambar 4.2 Satu Tweet Berisi Dua Jenis Informasi Berbeda

Hasil percobaan klasifikasi informasi kejadian dicatat dalam tabel confussion matrix seperti pada Tabel 4.15. Berdasarkan hasil pencatatan pada confussion matrix dapat dihitung nilai precision, recall, dan f-measure untuk

metode klasifikasi. Hasil perhitungan *precision*, *recall*, dan *f-measure* ditunjukkan pada Tabel 4.16.

Tabel 4.15 Tabel Confussion Matrix untuk Hasil Klasifikasi Informasi Kejadian

Kelas	Non-Informasi Kejadian Penting	Lalu - Lintas	Kebakaran	Bencana-Alam	Total
Non-Informasi Kejadian Penting	150	8	2	4	164
Lalu - Lintas	5	141	2	2	150
Kebakaran	1		49		50
Bencana-Alam	2			48	50
Total	158	149	53	54	

Tabel 4.16 Precision, Recall, Dan F-Measure untuk Hasil *NeuroNER*

Kelas	Prec (%)	Recall (%)	F1 (%)
Non-Informasi Kejadian Penting	91,46	94,94	93,17
Lalu - Lintas	94,00	94,63	94,31
Kebakaran	98,00	92,45	95,15
Bencana-Alam	96,00	88,89	92,31
Average	**94,87**	**92,73**	**93,73**

4.3.1.4 Ekstraksi Relasi dengan *NeuralRE*

Pada tahap ini, tweet yang sudah dikenali sebagai tweet informasi kejadian penting yaitu kelas Lalu-Lintas, Kebakaran, dan Bencana-Alam akan dilakukan proses ekstraksi relasi antar entitas jika terdapat **lebih dari satu entitas lokasi**. Namun, terlalu banyaknya entitas lokasi dalam satu tweet informasi kejadian akan cenderung sulit dipahami baik oleh sistem maupun pembaca, karena penyampaian yang tidak lugas dan kongkrit dalam penulisan informasi kejadian. Oleh karena itu, sistem hanya akan memproses tweet informasi kejadian penting yang mempunyai entitas lokasi tidak lebih dari empat. Contoh dari tweet berisi lebih dari empat entitas lokasi ditunjukkan pada Tabel 4.17.

Tabel 4.17 Contoh Tweet dengan Jumlah Entitas Lebih dari Empat

No	Teks Tweet	Jumlah Entitas Lokasi	Total Relasi
1	@e100ss **Taman** arah **Sby** padat dari Seblm **pertigaan Brimob Ketegan** sampai tanjakan yg mau masuk **Tol Gresik**,Merambat seblm **Bunderan Citto** dari arah **Mojokerto**,**Dolog** padat,**Jemursari** Ramai padat,Sekitar **plaza Marina** Ramai lancar	9	9! / 2!(9-2)! = **36**
2	@e100ss @PuspitaFM @LalinNews @infomalangraya @gen1031fm @infomalang Lalin **lawang - malang** merambat mulai **SMP 1 Singosari**. Arah ke **Lawang** macet krn ada truk tangki gandeng yg mogok setelah **Alam Hijau**. Lbh baik ambil jln alternatif ke **Lawang exit sumberwuni**.	6	6! / 2!(6-2)! = **15**

Hasil keluaran dari proses ekstraksi relasi dengan *NeuralRE* ditunjukkan pada Tabel 4.18 yang kemudian dicatat dalam tabel *confussion matrix* seperti pada Tabel 4.19. Berdasarkan hasil pencatatan pada *confussion matrix* dapat dihitung nilai *precision, recall,* dan *f-measure* untuk metode NeuralRE. Hasil perhitungan *precision, recall,* dan *f-measure* ditunjukkan pada Tabel 4.20.

Tabel 4.18 Hasil Ekstraksi Relasi dengan *NeuralRE*

No	Teks Tweet	Input NeuralRE	Hasil Ekstraksi Relasi
1	RT @WahyudiJo: @e100ss terjadi tabrak beruntun **km 6 800 tol satelit ps turi**, sepertinya mobil inova masih di kanan jalan	terjadi tabrak beruntun **<e1>hwymse<e1>** **<e2>loc<e2>** , sepertinya obj inova masih di kanan jalan	Highway-Position (e2,e1)
2	RT @ikaripanda: Antrian panjang **gate porong** arah **sidoarjo** @e100ss https://t.co/3sTHSNchhE	antrian panjang **<e1>loc<e1>** arah **<e2>gpe<e2>**	StartingPoint-Destination (e1,e2)

No	Teks Tweet	Input NeuralRE	Hasil Ekstraksi Relasi
3	10.15 : tol **Waru** - **Satelit** padat sejak **km 10.** dominasi kendaraan besar.;	• time : **<e1>loc<e1>** - **<e2>loc<e2>** padat sejak hwymse dominasi kendaraan besar . • time : **<e1>loc<e1>** - loc padat sejak **<e2>hwymse<e2>** dominasi kendaraan besar . • time : loc - **<e1>loc<e1>** padat sejak **<e2>hwymse<e2>** dominasi kendaraan besar	StartingPoint-Destination (e1,e2) Highway-Position (e1,e2) Other
4	**ahmad yani** depan **royal** 2 arah padat	**<e1>loc<e1>** depan **<e2>bld<e2>** 2 arah padat	Street-Place (e1,e2)
5	**jalan jaksa agung suprapto sukorejo lamongan** arah **surabaya** padat, imbas volume. indra via melaporkan, antrean sejak menjelang **rs muhammadiyah lamongan**	• **<e1>loc<e1>** arah **<e2>gpe<e2>** padat, imbas volume. indra via melaporkan, antrean sejak menjelang **bld** • **<e1>loc<e1>** arah **gpe** padat, imbas volume. indra via melaporkan, antrean sejak menjelang **<e2>bld<e2>** • **loc** arah **<e1>gpe<e1>** padat, imbas volume. indra via melaporkan, antrean sejak menjelang **<e2>bld<e2>**	StartingPoint-Destination (e1,e2) Street-Place (e1,e2) StartingPoint-Destination (e2,e1)

Tabel 4.19 Tabel Confussion Matrix untuk Hasil Ekstraksi Relasi (*NeuralRE*)

Relasi	Other	Higway-Pos (e1,e2)	Higway-Pos(e2,e1)	Street-Place (e1,e2)	Street-Place (e1,e2)2	Start-Dest (e1,e2)	Start-Dest (e2,e1)
Other	11	1				3	2
Higway-Pos (e1,e2)	0	8				2	
Higway-Pos (e1,e2)	0		3				
Street-Place (e1,e2)	0			19			
Street-Place (e1,e2)	0				8		
Start-Dest (e1,e2)	0			1		102	4
Start-Dest (e2,e1)	0						21

Tabel 4.20 Precision, Recall, Dan F-Measure untuk Hasil Ekstraksi Relasi (*NeuralRE*)

Relasi	**Prec (%)**	**Recall (%)**	**F1 (%)**
Other	64,71	100,00	78,57
Higway-Pos (e1,e2)	80,00	88,89	84,21
Higway-Pos (e1,e2)	100,00	100,00	100,00
Street-Place (e1,e2)	100,00	95,00	97,44
Street-Place (e1,e2)	100,00	100,00	100,00
Start-Dest (e1,e2)	95,33	95,33	95,33
Start-Dest (e2,e1)	100,00	77,78	87,50
Average	**91,43**	**93,86**	**91,86**

4.3.1.5 Standarisasi Nama Entitas

Kesalahan penulisan nama tempat dalam sebuah tweet tidak dapat dihindari. Boleh jadi sistem dapat melakukan pengenalan nama lokasi dengan benar meski penulisannya salah, namun dalam proses visualisasi nanti akan terjadi kendala geocoding oleh GoogleMaps. Oleh karena itu perlu dilakukan perbaikan

nama entitas lokasi dengan metode FASM. Beberapa tweet dengan kesalahan penulisan nama lokasi ditunjukkan pada Gambar 4.2.

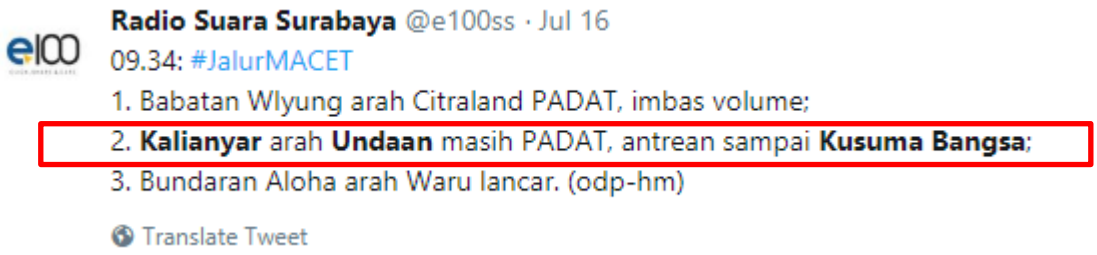
Standarisasi penulisan Gazetter bisa justru menjadi salah dalam memperbaiki nama lokasi karena adanya penggunaan singkatan yang kurang tepat dalam penulisan nama lokasi. Selain itu, penulisan nama tempat sudah benar namun karena tidak terdapat dalam kamus Gazetter sehingga setelah melalui proses standarisasi nama, nama lokasi justru menjadi tidak sesuai dengan nama yang dimaksud dalam tweet. Hal tersebut dapat dilihat pada Tabel 4.15.

Tabel 4.21 Contoh Kesalahan dalam Proses Standarisasi Nama Gazetter dengan FASM

Teks Tweet	**Hasil Standarisasi**
lahan tebu ada api kebakaran di km 698 tol sumo dari **jombang** ke mojokerto	ahan tebu ada api kebakaran di km 698 tol sumo dari **kembang** ke mojokerto
kalapas sukamiskin terang-terangan minta mahar mobil untuk kamar mewah napi (odp-pr)	**kapas suka skin** terang-terangan minta mahar mobil untuk kamar mewah napi (odp-pr)
arah **probolinggo** macet parah	arah **pro lingga** macet parah

4.3.1.6 Konversi Hasil Ekstraksi Informasi ke dalam Bentuk Graf

Dalam satu tweet jika mempunyai lebih dari dua entitas maka akan memiliki lebih dari satu relasi, maka perlu didapatkan keterhubungan antar relasi yang terbentuk. keterhubungan antar relasi bisa didapatkan dengan mudah dengan mengubahnya kebentuk directed-graf. Contoh tweet yang memiliki lebih dari satu relasi ditunjukkan pada Gambar 4.3. Pada gambar tersebut didapatkan hasil ekstraksi seperti pada Gambar 4.4. Penentuan startingpoint, destination, dan waypoints dari hasil ekstraksi dengan memanfaatkan graf seperti ditunjukkan pada Gambar 4.5.

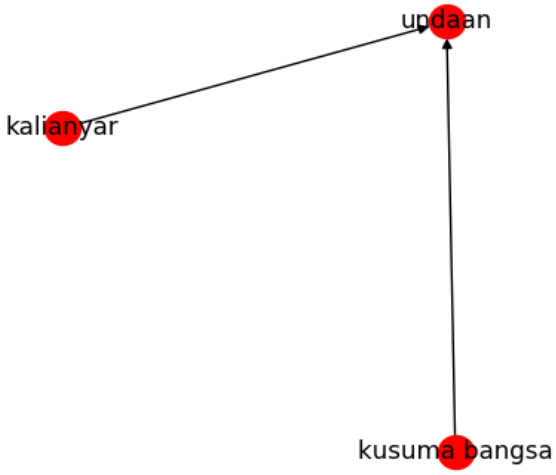


Gambar 4.3 Contoh Tweet dengan Entitas Lokasi lebih dari Dua

```
<e1>loc<e1> arah <e2>loc<e2> masih padat antrean sampai loc
<e1>loc<e1> arah loc masih padat antrean sampai <e2>loc<e2>
loc arah <e1>loc<e1> masih padat antrean sampai <e2>loc<e2>

[5. 3. 6.]
```

Gambar 4.4 Hasil keluaran Ekstraksi Relasi



Gambar 4.5 Hasil Konversi Hasil Ekstraksi Relasi dalam Bentuk Graf

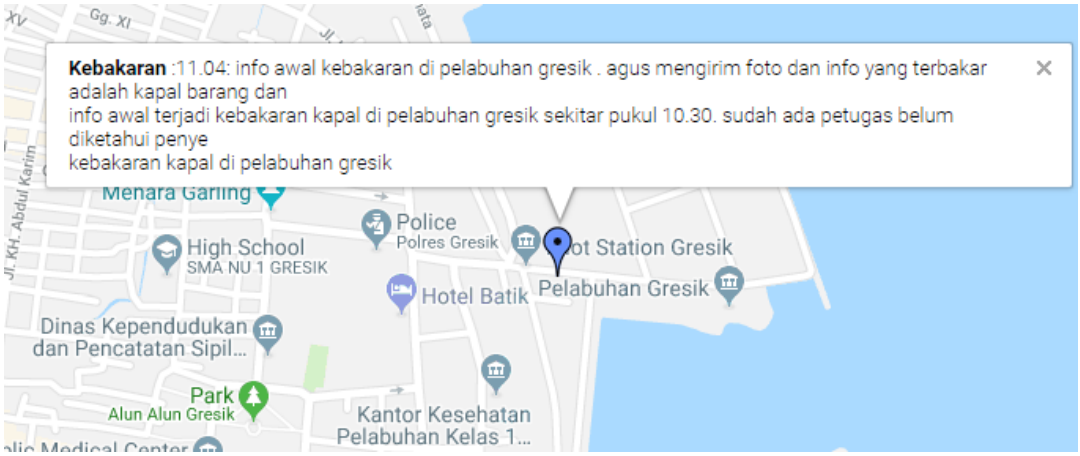
Gambar 4.6 Graf Hasil Penggabungan Node Kusuma Bangsa dan Kalianyar oleh Relasi Street-Place(e1,e2)

Dari Gambar 4.6 diterapkan perhitungan successor dan predecessor untuk menentukan node yang akan menjadi startingpoint dan destination. Berdasarkan hal tersebut dapat ditentukan bahwa kusuma bangsa kalianyar sebagai startingpoint dan undaaan sebagai destination.

4.3.1.7 Filter Deteksi Kesamaan dan Pengelompokan Informasi Kejadian

Untuk menghindari informasi dengan isi tweet yang sama dengan jumlah banyak yang dikarenakan banyaknya retweet dari suatu tweet, maka dilakukan filter kesamaan informasi kejadian. Sedangkan untuk kemudahan pembacaan informasi pada visualisasi data, untuk dua atau lebih informasi yang berbeda dan atau sama dengan kesamaan lokasi dan selisih waktu kejadian kurang dari 120 menit, maka dilakukan pengelompokan informasi kejadian berdasarkan kesamaan lokasi. Contoh hasil tahap ini ditunjukkan pada Gambar 4.7.

Pada Gambar 4.7 dapat dilihat bahwa dalam rentan waktu kurang dari 120 menit terdapat tiga tweet informasi kejadian penting dengan penulisan yang berbeda dengan isi informasi yang sama. Maka akan dikelompokan menjadi satu node dalam visualisasi yang selebihnya akan dibahas pada subbab berikutnya.

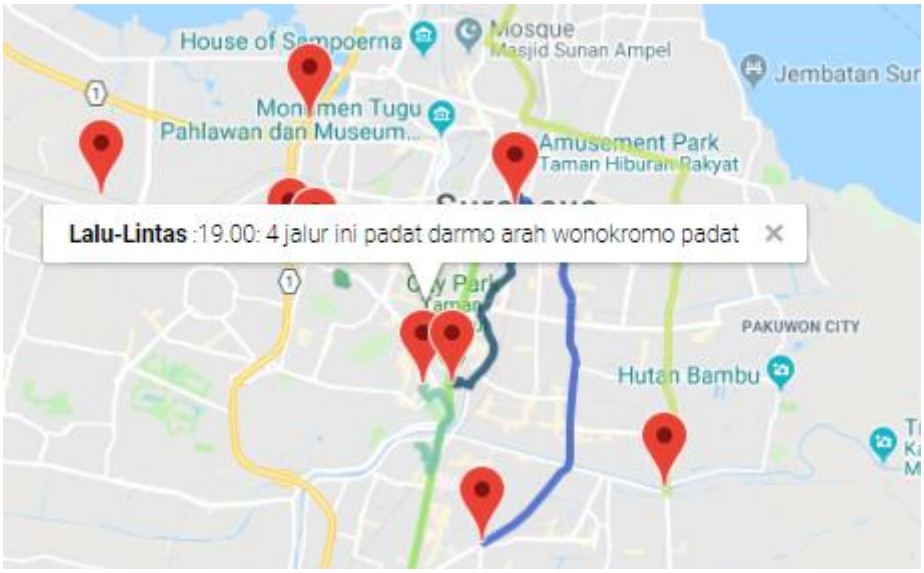


Gambar 4.7 Hasil Pengelompokan Informasi Kejadian Penting

4.3.1.8 Visualisasi Data

Visualisasi data menggunakan bantual tool API Google Maps V3 memudahkan khalayak ramai untuk mengetahui kejadian-kejadian penting yang ada di Indonesia khususnya wilayah Kota Surabaya dan Sidoarjo. Titik kejadian dan atau titik awal kejadian ditandai dengan marker. Jenis marker dibedakan

menajdi tiga yaitu merah untuk kejadian lalu-lintas, kuning untuk kejadian kebakaran, dan hijau untuk kejadian bencana-alam. Hasil visualisasi rentang waktu 120 menit bisa dilihat seperti pada Gambar 4.8.



Gambar 4.8 Hasil Visualisasi Data Informasi Kejadian Penting

4.3.2 Evaluasi Hasil Uji Coba Sistem Keseluruhan

Berdasarkan hasil ujicoba diatas, dilakukan perhitungan untuk mengevaluasi kinerja sistem secara keseluruhan dengan rumus persamaan Sensitivity. Perhitungan ini berdasarkan keberhasilan sistem melakukan ekstraksi informasi dari tahap awal hingga visualisasi data. Dengan TP adalah jumlah keberhasilan sistem dalam mengenali entitas, klasifikasi kejadian, dan ekstraksi relasi, maka jika dari ketiga proses tersebut tidak terdapat kesalahan, maka jumlah TP ditambah 1. Untuk FN adalah jumlah ketidak berhasilan sistem dalam mengenali entitas, klasifikasi kejadian, dan atau ekstraksi relasi, maka jika dalam ketiga proses tersebut terdapat kesalahan, maka jumlah FN ditambah 1.

Dari total data uji coba sebanyak 391, jumlah TP sebanyak 343 dan jumlah FN sebanyak 48. Berdasarkan rumus *sensitivity* dapat dihitung tingkat keberhasilan sistem dalam menjalankan tugasnya yaitu sebesar 87,72%. Jika dilihat dengan seksama, nilai sensitivity sistem secara keseluran lebih rendah daripada nilai rata-rata *f-measure* dari masing-masing proses. Hal ini dikarenakan dalam setiap satu alur pemrosesan yaitu satu tweet masuk ke dalam sistem kemudian diproses hingga

divisualisasikan secara benar oleh sistem atau dibuang karena karena tweet termasuk informasi tidak penting, terjadi kesalahan proses secara acak dan berbeda. Terkadang dalam satu alur proses terjadi kesalahan dalam pengenalan entitas bernama saja, berikutnya terjadi kesalahan pada klasifikasi jenis kejadian, dan atau terjadi pada proses keduanya. Sehingga persentase *false negatif* pada sistem secara keseluruhan lebih besar dari pada persentase *false negatif* pada masing-masing tahapan proses.

BAB 5
Penutup

Pada bab terakhir pada buku thesis ini, dapat diambil kesimpulan dari hasil percobaan pada bab 4 untuk didapatkan sara-saran guna kemungkinan pengembangan untuk penelitian berikutnya berdasarkan saran yang ada.

5.1 Kesimpulan

Berdasarkan implementasi serangkaian uji coba yang telah dilakukan dengan aplikasi ekstraksi informasi kejadian penting pada data stream twitter, maka dapat ditarik beberapa kesimpulan dari penelitian ini sebagai berikut:

1) Pengenalan entitas bernama menggunakan model *NeuroNER* memiliki performa mengenali entitas bernama yang cukup baik yaitu dengan nilai precision, recall, dan f-measure masing-masing 96,56%, 95,89%, dan 96,21%
2) Klasifikasi jenis kejadian informasi menggunakan RCNN dapat melakukan mengenali tweet sesuai jenisnya dengan baik yang ditunjukkan nilai precision, recall, dan f-measure masing-masing 94,87%, 92,73%, dan 93,73%
3) Ekstraksi relasi antar entitas bernama bekerja cukup baik meski tidak sebaik performa dari proses lainnya yaitu dengan nilai precision, recall, dan f-measure masing-masing 91,43%, 93,86%, dan 91,86%
4) Standarisasi nama entitas Gazetter dapat melakukan perbaikan dengan tingkat kebenaran 100% tidak termasuk nama Gazetter luar kota Surabaya dan Sidoarjo.
5) Tingkat keberhasilan sistem dari tahap *streaming* data hingga visualisasi data dengan menggunakan yang diusulkan mencapai 87,72%.
6) Sistem yang diusulkan dapat melakukan ekstraksi informasi secara (mendekati) waktu nyata dengan waktu komputasi yang tercepat 0.002 detik untuk satu tweet dan dengan rata-rata 1.2 detik untuk satu tweet.

5.2 Saran

Dari analisan hasil dan kesimpulan maka didapat beberapa saran untuk pengembangan sistem lebih lanjut antara lain:

1) Sistem ekstraksi informasi kejadian penting masih pada lingkup area kota Surabaya dan Sidoarjo. Diharapkan pada penelitian berikutnya dapat diperluas lingkup area penelitian mencakup provinsi atau bahkan nasional dengan menggunakan data latih berupa *tweet* dan data gazetter yang lebih banyak.
2) Untuk klasifikasi jenis kejadian penting diperlukan klasifikasi multi-label karena sebuah tweet dimungkinkan memiliki lebih dari satu jenis kejadian.
3) Untuk pemrosesan tweet bisa dilakukan dengan model antrian menggunakan *sliding window*, sehingga tweet tidak diproses satu per-satu setiap terdeteksi ada tweet baru, melainkan diproses dalam kurun waktu tertentu namun tidak melupakan pemrosesan (mendekati) waktu nyata, misalnya sekitar 5-10 detik. Hal ini diperlukan untuk menghindari tidak terprosesnya tweet yang diposting secara bersamaan dengan jumlah banyak.

DAFTAR PUSTAKA

[1] S. Kruikemeier, "Computers in Human Behavior How political candidates use Twitter and the impact on votes," *Comput. Human Behav.*, vol. 34, pp. 131–139, 2014.

[2] K. He, Y. Li, S. Soundarajan, and J. E. Hopcroft, "Hidden community detection in social networks," vol. 425, pp. 92–106, 2018.

[3] S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection in tweets streamed in real time : a big data approach," *Digit. Commun. Networks*, vol. 2, no. 3, pp. 108–121, 2016.

[4] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[5] J. Pal and A. Gonawela, "Studying political communication on Twitter: the case for small data," *Curr. Opin. Behav. Sci.*, vol. 18, pp. 97–102, 2017.

[6] P. Panagiotopoulos, J. Barnett, A. Z. Bigdeli, and S. Sams, "Social media in emergency management: Twitter as a tool for communicating risks to the public," *Technol. Forecast. Soc. Change*, vol. 111, pp. 86–96, 2016.

[7] J. Li *et al.*, "Social Media: New Perspectives to Improve Remote Sensing for Emergency Response," *Proc. IEEE*, vol. 105, no. 10, pp. 1900–1912, 2017.

[8] Y. Gu, Z. Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," *Transp. Res. Part C Emerg. Technol.*, vol. 67, pp. 321–342, 2016.

[9] M. Hasby, M. L. Khodra, and A. Purwarianti, "Optimal Path Finding based on Traffic Information Extraction from Twitter," in *Prosiding International COnference on ICT for Smart Smart Society*, 2013.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," 2016.

[11] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks," 2017.

[12] P. Gupta, C. Technology, B. Andrassy, and C. Technology, "Table Filling Multi-Task Recurrent Neural Network for Joint Entity and Relation Extraction," pp. 2537–2547, 2016.

[13] M. Miwa and M. Bansal, "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures," 2015.

[14] X. Jiang, Q. Wang, P. Li, and B. Wang, "Relation Extraction with Multi-

instance Multi-label Convolutional Neural Networks," no. 89, pp. 1471–1480, 2016.

[15] W. Garbe, "Fast approximate string matching with large edit distances in Big Data," 2015. [Online]. Available: http://blog.faroo.com/2015/03/24/fast-approximate-string-matching-with-large-edit-distances/. [Accessed: 25-Mar-2018].

[16] L. Derczynski *et al.*, "Analysis of named entity recognition and linking for tweets," *Inf. Process. Manag.*, vol. 51, no. 2, pp. 32–49, 2015.

[17] Z. Zhao *et al.*, "Disease named entity recognition from biomedical literature using a novel convolutional neural network," *BMC Med. Genomics*, vol. 10, no. Suppl 5, 2017.

[18] Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, "Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks," *Database*, vol. 2016, pp. 1–8, 2016.

[19] K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv*, 2015.

[20] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," *Twenty-Ninth AAAI Conf. Artif. Intell.*, pp. 2267–2273, 2015.

[21] A. Hagberg, D. Schult, and P. Swart, "NetworkX Reference," 2018.

RIWAYAT PENULIS

Fatra Nonggala Putra, Dilahirkan di Kabupaten Blitar tepatnya di Kelurahan Dandong Kecamatan Srengat pada tanggal 01 Nopember 1990. Peneliti menyelesaikan pendidikan Sekolah Dasar di SDN Dandong 1 di Kelurahan Dandong Kabupaten Blitar pada tahun tahun 2003. Pada tahun itu juga peneliti melanjutkan Pendidikan di SMP Negeri 2 Srengat Kecamatan Srengat dan tamat pada tahun 2006 kemudian melanjutkan Sekolah Menengah Atas di SMA Negeri 1 Srengat pada tahun 2006 dan lulus pada tahun 2009. Pada tahun 2009 peneliti melanjutkan pendidikan di perguruan tinggi negeri, tepatnya di Universitas Negeri Malang (UM), Jurusan Teknik Elektro Fakultas Teknik pada Program Studi Pendidikan Teknik Informatika (PTI). Peneliti menyelesaikan kuliah strata satu (S1) pada tahun 2014. Pada tahun 2016 peneliti melanjutkan pendidikan strata dua (S2) di Institut Teknologi Sepuluh Nopember Surabaya (ITS). Semasa kuliah s1 hinggga s2 penulis aktif di organisasi ekstra kampus yaitu Himpunan Mahasiswa Islam (HMI) dan sekarang aktif sebagai pengurus Badan Koordinasi HMI Jawa Timur periode 2016-2018.