

**PROPOSAL TESIS**

**Transfer Learning for Automatic Scoring Short Answer: A siamse Bidirectional LSTM Network Architecture**

**Agung Wiratmo**

**NRP. 05111850010026**

**DOSEN PEMBIMBING**

**Dr. Eng. Chastine Fatichah, M.Kom**

**NIP: 19751220 20011220 02**

**PROGRAM MAGISTER**

**DEPARTEMEN INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**SURABAYA**

**2018**

**LEMBAR PENGESAHAN** **PROPOSAL TESIS**

Judul :

Oleh : Agung Wiratmo

NRP : 05111850010026

Telah diseminarkan pada:

Tanggal : -

Tempat : -

Mengetahui / menyetujui

Dosen Penguji: Dosen Pembimbing:

1. 1.

Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

NIP NIP. 19751220 20011220 02

2.

NIP.

3.

.

NIP

[*Halaman ini sengaja dikosongkan*]

**Transfer Learning for Automatic Scoring Short Answer: A siamse Bidirectional LSTM Network Architecture**

Nama Mahasiswa : Agung Wiratmo

NRP : 05111850010026

Pembimbing I : Dr. Eng. Chastine Fatichah, S.Kom,

M.Kom

Pembimbing II : Chastine Fatichah.

# **ABSTRAK**

**Kata kunci** :

[*Halaman ini sengaja dikosongkan*]

Name : Agung Wiratmo

Student Identity Number : 05111850010026

Supervisor : Dr. Eng. Chastine Fatichah, S.Kom,

M.Kom

# **ABSTRACT**

**Keywords** :

# **DAFTAR ISI**

[**ABSTRAK** v](#_Toc515261449)

[**ABSTRACT** vii](#_Toc515261450)

[**DAFTAR ISI** ix](#_Toc515261451)

[**DAFTAR GAMBAR** xii](#_Toc515261452)

[**DAFTAR TABEL** xiv](#_Toc515261453)

[**BAB 1 PENDAHULUAN** 1](#_Toc515261454)

[1.1. Latar Belakang 1](#_Toc515261455)

[1.2. Perumusan Masalah 4](#_Toc515261456)

[1.3. Tujuan 4](#_Toc515261457)

[1.4. Manfaat 4](#_Toc515261458)

[1.5. Kontribusi Penelitian 5](#_Toc515261459)

[1.6. Batasan Masalah 5](#_Toc515261460)

[**BAB 2 KAJIAN PUSTAKA** 7](#_Toc515261461)

[2.1 Informasi Kejadian Penting 7](#_Toc515261462)

[2.2 Ekstraksi Informasi 7](#_Toc515261463)

[2.3 Neuro Named Entity Recognition (*NeuroNER*) 8](#_Toc515261464)

[2.4 *Relation Extraction* (RE) 8](#_Toc515261465)

[2.5 Support Vector Machine (SVM) 9](#_Toc515261466)

[2.6 Fast Approximate String Matching 9](#_Toc515261467)

[2.7 N-Gram 10](#_Toc515261468)

[2.8 Penelitian yang Relavan 11](#_Toc515261469)

[**BAB 3 METODOLOGI PENELITIAN** 13](#_Toc515261470)

[3.1 Studi Literatur 13](#_Toc515261471)

[3.2 Sistem Deteksi Kejadian 14](#_Toc515261472)

[3.3 Pengambilan Data 16](#_Toc515261473)

[3.3.1 Praproses 17](#_Toc515261474)

[3.3.2 Ekstraksi Fitur 18](#_Toc515261475)

[3.3.3 Klasifikasi Tweet kejadian Penting 18](#_Toc515261476)

[3.3.4 Ekstraksi informasi berbasis *NeuroNER* 19](#_Toc515261477)

[3.3.5 Ekstrasi Informasi Berbasis *Relation Extraction* (RE) 20](#_Toc515261478)

[3.3.6 Standarisasi Entitas Nama 21](#_Toc515261479)

[3.3.7 Klasifikasi Jenis Kejadian 22](#_Toc515261480)

[3.3.8 Visualisasi Data 22](#_Toc515261481)

[3.4 Ujicoba dan Analisa Hasil 22](#_Toc515261482)

[3.4.1 Akurasi 23](#_Toc515261483)

[3.4.2 Precision, Recall, dan F-Measure 23](#_Toc515261484)

[3.5 Dokumentasi dan Jadwal Penelitian 23](#_Toc515261485)

[**DAFTAR PUSTAKA** 25](#_Toc515261486)

# **DAFTAR GAMBAR**

[Gambar 2.1 Arsitektur *Neuroner* 8](#_Toc514058292)

[Gambar 2.2 Proses Pencarian Hyperplane Terbaik Untuk Memisahkan Dua Kelas 9](#_Toc514058293)

[Gambar 3.1 Diagram Alur Metodologi Penelitian 13](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058294)

[Gambar 3.2 Desain Sistem Deteksi Kejadian 15](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058295)

[Gambar 3.3 Pembentukan Daftar Persamaan Kata Tidak Baku Pada Twitter 16](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058296)

[Gambar 3.4 Diagram Alur Pembentukan Gazetteer 16](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058297)

[Gambar 3.5 Diagram Alur Tahapan Praproses, Ekstraksi Fitur, Dan Klasifikasi 18](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058298)

[Gambar 3.6 Contoh Pelabelan Data Latih Neuroner 19](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058299)

[Gambar 3.7 Diagram Alur Pelatihan Neuroner 20](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058300)

[Gambar 3.8 Diagram Alur Ekstraksi Entitas Bernama Dengan Neuroner 20](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058301)

[Gambar 3.9 Diagram Alur Ekstraksi Relasi Antar Entitas Bernama 21](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058302)

[Gambar 3.10 Diagram Alur Standarisasi Entitas Nama 21](file:///D:\Kuliah%20S2\Thesis\LAporan\Proposal%20REVISI%20MEI.docx#_Toc514058303)

[*Halaman ini sengaja dikosongkan*]

# **DAFTAR TABEL**

[tabel 2.1 Contoh Pencarian Perbaikan String dengan Fasm 10](#_Toc512512802)

[Tabel 2.2 Contoh Word N-Gram Dengan Model *Uni-Gram* dan *Bi-Gram* 10](#_Toc512512803)

[Tabel 2.3 Contoh Letter N-Gram Dengan Model *Bi-Gram* dan *Tri-Gram* 10](#_Toc512512804)

[Tabel 2.4 Penellitian yang Relevan 11](#_Toc512512805)

[Tabel 3.1contoh Daftar Kata Tidak Baku Hasil Filter dari Tweet 17](#_Toc512512806)

[Tabel 3.2 Contoh Gazetter Nama Lokasi dan Tempat 17](#_Toc512512807)

[Tabel 3.3 Daftar Notasi Entitas Bernama 19](#_Toc512512808)

[Tabel 3.4 Contoh Ekstraksi Relasi Antar Entitas dengan RE 21](#_Toc512512809)

[Tabel 3.5 Jadwal Kegiatan Penelitian 24](#_Toc512512810)

[*Halaman ini sengaja dikosongka]*

# **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. Latar Belakang

Penilaian hasil belajar yang dilakukan oleh pendidik bertujuan mengevaluasi, memantau proses pembelajaran, serta perbaikan hasil belajar dilakukan secara berkesinambungan­­ ­(Kemendikbud, 2016). Teknik penilaian yang digunakan adalah tes tertulis, tes lisan, dan penugasan. Tes tertulis adalah tes yang soal dan jawaban disajikan secara tertulis berupa pilihan ganda, isian, benar-salah, menjodohkan, dan uraian (Tim Direktorat Pembinaan SMP, 2017). Dalam penilaian hasil belajar terdaat beberapa prinsip yang harus diperhatikan yaitu sahih, objektif, adil, terpadu, terbuka, menyeluruh, sistematis, beracuan kriteria, akuntabel (Kemendikbud, 2016). Penilaian pada tes tertulis dengan jenis uraian memakan waktu yang lebih lama daripada penilaian pilihan ganda serta memiliki kecenderungan subjektif lebih besar dibandingkan dengan tes yang lainya.

Penilaian hasil belajar dengan bentuk uraian banyak menjadi objek penelitian dalam beberapa tahun terakhir. Salah satu tujuan utamanya adalah membuat penilaian secara otomatis dengan berbagai metode yang dilakukan oleh beberapa peneliti. Lahitani melakukan pengoptimalan pada metode TF-IDF yang akan memiliki nilai kemiripan yang kecil jika perbandingan antar dokumen tidak ada yang sama walaupun memiiki kedekatan makna (Lahitani, 2017). Dalam penelitiannya, Lahitani melakukan pengujian kemiripan dokumen dengan mengkombinasikan pembobotan kata TF-IDF-DF dengan pendekatan *Cosine Similarity*.

Liang et al melakukan pengimplementasian penilaian pada soal esai dengan mempertimbangkan kriteria yang telah ditetapkan oleh expert (Liang *et al.*, 2018). Penelitian Liang dkk menggunakan metode

1. 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 SVM?
3. Bagaimana cara melakukan ekstraksi informasi pada data Twitter menggunakan metode *NeuroNER*?
4. Bagaimana mengetahui relasi antar entitas menggunakan metode RE?
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?
7. Tujuan

Tujuan yang akan dicapai dalam pembuatan tesis ini adalah deteksi kejadian pada data stream twitter dengan melakukan perbaikan ekstraksi informasi menggunakan *NeuroNER*, RE, dan FASM.

1. Manfaat

Manfaat dari penelitian ini adalah:

1. Mengetahui praproses yang sesuai untuk data tweet yang digunakan pada proses ekstraksi informasi kejadian
2. Menghasilkan metode terbaik untuk melakukan ekstraksi informasi pada data stream Twitter khususnya ekstraksi entitas nama lokasi.
3. Memberikan informasi secara cepat dan utuh kepada masyarakat dengan melakukan pemetaan dan pengelompokan informasi kejadian berdasarkan lokasi kejadian secara akurat melalui peta digital.
4. Ekstraksi informasi secara sistematis dan tersimpan dalam sistem dapat digunakan pihak terkait untuk mengetahui 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.
5. Kontribusi Penelitian

Kontribusi pada penelitian ini adalah penggunaan *NeuroNER*, RE, 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 RE serta penggunaan metode *FASM* untuk standarisasi entitas nama lokasi yang telah dikenali oleh NeuroNER.

1. 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 pengeumpulan 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 seputardeteksi kejadian, ekstrasi informasi, dan klasifikasi*.*

* 1. 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.

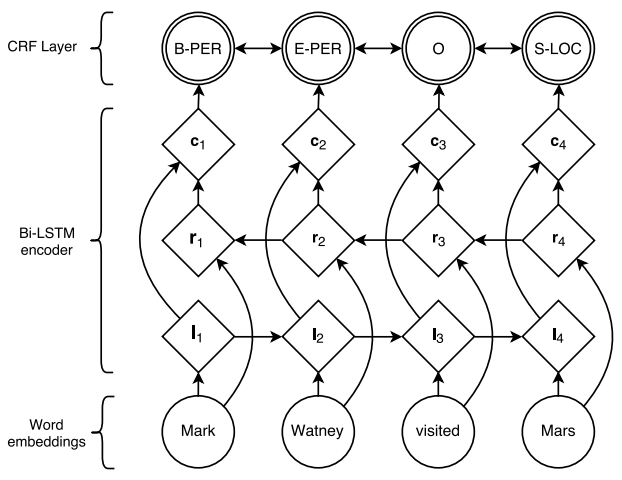
Berdasarkan [5] dan definisi di atas, maka dapat diformulasikan kembali karakteristik dari suatu informasi termasuk kedalam informasi kejadian penting adalah memiliki 3 unsur: 1) waktu: informasi berupa kejadian dengan waktu pemberitaan mendekati waktu-nyata (*real-tim*e). Jadi, selisih waktu kejadian 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 berdampak negatif bagi masyarakat dalam beraktifitas haruslah memiliki entitas lokasi. Tanpa adanya entitas lokasi di dalam informasi tersebut, maka informasi menjadi tidak penting.

* 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) [18]. 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 rekam medis [13]. Dalam deteksi kejadian, NER digunakan untuk mengenali entitas nama lokasi dan tempat [6].

* 1. 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) [7]. *NeuroNER* membuat alur prediksi anotasi berjalan lebih baik dengan ANNs dan CRF daripada hanya menggunakan CRF [9] . Arsitektur dari *NeuroNer* ditunjukkan pada Gambar 2.1.



Gambar . Arsitektur *NeuroNER* [7]

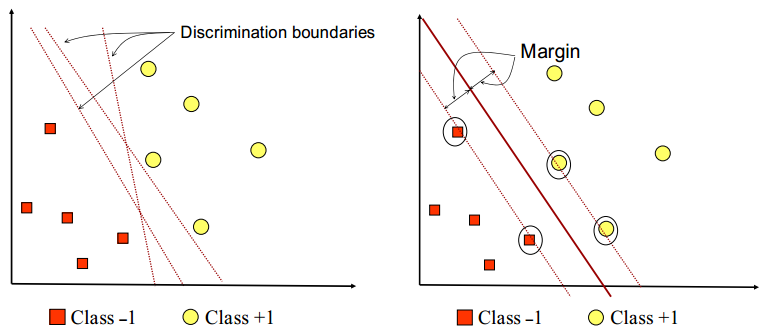
* 1. *Relation Extraction* (RE)

RE adalah metode ekstraksi informasi yang digunakan untuk mengetahui relasi antar entitas berdasarkan jenis entitas yang dihubungkan dengan relasi [15][16]. Ada berbagai macam model RE [15] namun, pada penelitian ini model RE yang digunakan adalah RE dengan pencarian pola (RE *search for pattern*). Pada [6] mendefinisikan beberapa kata yang dijadikan pengenal relasi antara lain, ‘arah’, ‘sampai’, ‘menuju’, ‘*exit*’, ‘sekitar’, dan ‘kawasan’. Sedangkan untuk tanda baca yang didefinisikan sebagai relasi adalah tanda strip ‘ - ‘ yang mempunya arti sama seperti sampai / arah / antara.

* 1. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik [17]. Gambar 2.2 menunjukkan proses klasifikasi SVM untuk menemukan batas pemisah antar kelas (*Hyperplane*) terbaik untuk memisahkan dua kelas.

.Algoritma klasifikasi SVM digunakan untuk melakukan klasifikasi data Twitter kedalam beberapa kelas yang berbeda. Penggunaan *machine learning* tersebut dikarenakan performansinya yang handal dalam memprediksi kelas dari suatu data baru.



Gambar . Proses pencarian hyperplane terbaik untuk memisahkan dua kelas

* 1. 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) [19]. FASM memiliki performa prediksi dan waktu komputasi yang lebih baik daripada fuzzy string matching standar. Contoh dari FASM ditunjukkan dalam tabel 2.1.

Tabel . 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 |

* 1. N-Gram

Ada dua jenis N-gram yaitu word n-gram dan letter n-gram. Word n-gram adalah teknik memecah string menjadi potongan n kata. Contoh dari word n-gram dengan model uni-gram dan bi-gram ditunjukkan pada tabel 2.2. Sedangkan letter n-gram merupakan teknik memecah string menjadi beberapa potongan huruf atau karakter. Contoh dari letter n-gram dengan model bi-gram dan tri-gram ditunjukkan pada tabel 2.3.

Tabel . Contoh word n-gram dengan model *uni-gram* dan *bi-gram*

|  |  |  |
| --- | --- | --- |
| **String** | **uni-gram** | **bi-gram** |
| ekstraksi informasi kejadian | ektsraksi,  informasi,  kejadian | \_ ekstrasksi,  ekstraksi informasi,  informasi kejadian,  kejadian \_ |

Tabel . Contoh letter n-gram dengan model *bi-gram* dan *tri-gram*

|  |  |  |
| --- | --- | --- |
| **String** | **bi-gram** | **tri-gram** |
| informasi | \_i, in, nf, fo, or,  rm, ma, as, si, i\_ | \_in, inf, nfo, for, orm,  rma, mas, asi, si\_ |

* 1. 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. Pada [6] yang membandingkan antara NER dan RE untuk ekstraksi lokasi maka pada penelitian ini memilih untuk menggunakan keduanya yaitu RE sebagai pelengkap NER untuk mendapatkan ekstraksi entitas bernama dan ekstraksi relasi antar entitas. Model NER yang digunakan pada penelitian ini mengadopsi metode *NeuroNER* [9][14] 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 . Penellitian yang Relevan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NNo** | **Referensi** | **Deskripsi** | **Kelebihan** | **Kekurangan** |
| 11 | (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 NaiveBayes * Membedakan lokasi kejadian di jalan umum dan jalan toll dengan template khusus | * Bergantung pada software berbayar arcGIS untuk mengesktrak geocode lokasi. * Jika terjadi kesalahan penulisan lokasi maka sistem tidak dapat melakukan geocode. |
| 2  2 | (M. Hasby, 2013) | * Membandingkan dua metode yaitu NER (named entity recognition) dan RE (relation extration) untuk ekstraksi informasi dari tweet. | * Menggunakan NER untuk men-genali entitas lokasi | * Tidak menjelaskan berapa banyak data tweet yang digunakan untuk training dan testing. * Tidak memperhitungkan relasi antar lokasi |
|  |  |  |  |  |
| **NNo** | **Referensi** | **Deskripsi** | **Kelebihan** | **Kekurangan** |
| 33 | (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 |
| 44 | (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 **Error! Reference source not found.**.

Studi Literatur

Pengambilan Data

Dokumentasi Penulisan Laporan

Ujicoba dan Analisa Hasil

Desain Sistem

Gambar . Diagram Alur Metodologi Penelitian

Penjelasan tahapan metode penelitian pada **Error! Reference source not found.** akan diterangkan secara terperinci pada subbab berikut.

* 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 tekskhususnya ekstraksi informasi, NER, RE, dan algoritma klasifikasi. Berdasarkan studi literatur yang telah dilakukan, dapat diambil informasi sebagai berikut:

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

Deteksi kejadian diawali dengan melakukan streaming data pada akun twitter @e100ss. Setiap ada data tweet baru akan digunakan sebagai data input yang kemudian dilakukan tahapan praproses untuk membuat data menjadi layak. Dari tahapan praproses data diakukan klasifikasi untuk mendeteksi tweet yang masuk tergolong tweet kejadian penting atau bukan. Jika tweet teridentifikasi sebagai tweet kejadian penting maka akan dilakukan proses selanjutnya seperti pada P1 Gambar 3.4. Gambar 3.4 merupakan diagram alur sistem secara keseluruhan.

Gazetter

Klasifikasi Jenis Kejadian

(SVM)

Twitter

Crawling

Praproses

OSM

Parsing

**Ekstraksi Informasi**

**(NeuroNER, RE, FASM)**

Twitter

Stream

Ekstraksi Fitur

Praproses

**Visualisasi Data**

Data Tweet dengan notasi BIO

: Stream data Twitter

(Fase *Testing*)

: crawling data Twitter

(Fase *Training*)

: pebentukan Gazetter

Gambar . Desain Sistem Deteksi Kejadian

Klasifikasi tweet kejadian dan non-kejadian

(SVM)

* 1. Pengambilan Data

Ada tiga jenis data yang dibutuhkan dalam penelitian ini yaitu 1) data Tweet, 2) daftar kata tidak baku populer yang sering digunakan dalam Twitter, dan 3) 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.

Pertama, Pengumpulan data tweet dilakukan dengan melakukan *crawling* data *tweet* menggunakan API Twitter yang ditujukan untuk akun Radio Suara Surabaya (@e100ss) dan *crwaling* berdasarkan kata kunci yang dilakukan pada bulan April-Mei.

**Tokenizing**

**Crawling Twitter**

**KBBI**

**Filter Kata Tidak Baku**

**Daftar Kata Tidak Baku**

**Token** yang tidak terdaftar pada KBBI

Gambar . Pembentukan Daftar Persamaan Kata Tidak Baku pada Twitter

Kedua, membuat daftar kata tidak baku yang populer dari *tweet* bahasa indonesia dengan melakukan filter kata tidak baku terhadap token tweet dengan menggunakan kamus online bahasa indonesia. Kata yang tidak baku kemudian dibuatkan persamaan kata baku berdasarkan KBBI dan kemudian disimpan pada *database*. Alur pembentukan daftar kata khusus dan tidak baku seperti pada Gambar 3.3 dan contoh daftar kata tidak baku seperti pada tabel 3.1.

Gambar . Diagram Alur Pembentukan Gazetteer

**Parsing Data**

**Openstreetmap.org (OSM)**

**DB Gazetteer**

**Gazetteer**

Tabel . Contoh Daftar Kata Tidak Baku Hasil Filter dari Tweet

|  |  |
| --- | --- |
| **Kata Tidak Baku** | **Kata Baku** |
| bund  per3an  tl  pdt | bundaran  pertigaan  traffic light  padat |

Ketiga, 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.

Tabel . Contoh Gazetter Nama Lokasi dan Tempat

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

* + 1. 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) Remove url, mention, and hashtag: menghapus substring seperti mention dan link juga karakter ‘#’ pada data tweet. 3) normalisasi: mengubah kata singkatan kebentuk panjangnya dan kata tidak baku menjadi bentuk kata baku sesuai KBBI.

Data Tweet

**Daftar kata tidak baku**

Casefolding

Normalisasi

Ekstraksi Fitur (bi-gram)

Jenis Tweet

Klasifikasi (SVM)

(S

**Praproses**

Gambar . Diagram Alur Tahapan Praproses, Ekstraksi Fitur, dan Klasifikasi

* + 1. Ekstraksi Fitur

Ekstrasi fitur Data yang telah melalui tahap praproses selanjutnya akan dilakukan ekstraksi fitur menggunakan model *bigram* untuk mendapatkan fitur yang akan digunakan sebagai fitur model. Kumpulan kata hasil ekstrasi fitur akan digunakan sebagai penciri dari kelas tweet kejadian penting dan kelas tweet bukan kejadian penting, serta empat kelas jenis kejadian penting. Pada [12]][13] menunjukkan bahwa penggunaan fitur ekstraksi model *bi-gram* pada klasifier SVM dapat meningkatkan performa dan hasil klasifikasi. Diagram alur ekstraksi fitur ditunjukkan pada gambar 3.5.

* + 1. Klasifikasi Tweet kejadian Penting

Tahapan berikutnya setelah praproses adalah deteksi jenis tweet kejadian penting menggunakan klasifikasi SVM. Tweet diklasifikasikan menjadi 2 kelas yaitu 1) tweet kejadian penting, dan 2) tweet non-kejadian penting. Tweet yang termasuk kedalam kelas kejadian penting akan diteruskan ke tahapan proses selanjutnya. Sedangkan tweet yang masuk kedalam kelas non-kejadian penting akan diabaikan atau tidak diteruskan ke tahapan proses selanjutnya.

* + 1. 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) PRD (Product), 4) BLD (Building), 5) NPL (Natural Place), dan 6) MSE(Measurement) seperti pada tabel 3.1. Agar sistem dapat melakukan ektraksi 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. B (Begin) jika entitas merupakan kata pertama dari entitas, I (Inside) jika kata bukan merupakan kata pertama dari nama entitas, O (Outside), jika kata tidak termasuk dalam kategori entitas apapun. Contoh dari pelabelan data ditunjukkan pada gambar 3.6.

Tabel . Daftar notasi entitas bernama

|  |  |
| --- | --- |
| LOC (Location) | Wonokromo, Mulyorejo |
| GPE (Geograpical Political Entity) | Surabaya, Sidoarjo |
| PRD (Product) | Innova, avanza |
| BLD (Building) | Swiss-Belinn, Tunjungan Plaza |
| NPL (Natural Place) | Sungai jagir, Gunung Bromo |
| MSE (Measurement) | KM 90, 5.6 SR, 40 cm |

O

B-LOC

B-LOC

**Gate** **brebek** arah **waru** macet, gate buka cuman 1 sblah kiri

I-LOC

Gambar . 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*. Alur pelatihan *NeuroNER* ditunjukkan pada gambar 3.7.

Wonokromo arah A. Yani macet parah

Galaxy Mall I

B-LOC

B-LOC

B-BLD

**Validation Set**

kertajaya menuju gubeng padat merayap

pickup terjebur sungai jagir di depan kantor pajak

B-LOC

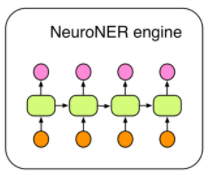
B-LOC

B-OBJ

B-NPL

B-BLD

**Train Set**



**Word2Vec Wikipedia**

**Gazetteer**

E m b e d d i n g

Gambar . Diagram Alur Pelatihan NeuroNER

NeuroNER

Data Tweet Setelah praproses

Entitas Bernama yang dikenali

Gambar . Diagram Alur Ekstraksi Entitas Bernama dengan NeuroNER

* + 1. Ekstrasi Informasi Berbasis *Relation Extraction* (RE)

Fungsi dari RE pada penelitian ini digunakan untuk melakukan ekstraksi relasi antar entitas yang sudah dikenali oleh NER. Hasil ekstraksi relasi berupa label berguna untuk mengetahui hubungan entitas satu dengan entitas lainnya yang dikenali oleh NER. Diagram alur RE ditunjukkan pada gambar 3.9 dan hasil ekstraksi pada tabel 3.4

Gambar . Diagram Alur Ekstraksi Relasi antar Entitas Bernama

Entitas Bernama yang dikenali

Relasi antar Entitas

RE I

Tabel . Contoh Ekstraksi Relasi antar Entitas dengan RE

|  |  |  |  |
| --- | --- | --- | --- |
| **Tweet: “**Manyar **arah** kertaya macet parah ekor **sampai** swiss bellin” | | | |
| **Entity Mention** | **Boundaries** | | **Entity Type** |
| manyar | (0,0) | | LOC |
| kertajaya | (2,2) | | LOC |
| swiss bellin | (7,8) | | BLD |
|  | | | |
| **Entity Mention Relation** | | **Relation Type** | |
| manyar, kertajaya | | StartingPoint - Destination | |
| manyar, swiss bellin | | StartingCongestion - EndCongestion | |

* + 1. 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.

**Standarisasi Entitas Nama**

**(FASM)**

Gazetteer

**Standar Entitas Nama**

“Jalan Ahmad Yani”

**Entitas yang Berhasil di Ekstrak**

“A Yani”

Gambar . Diagram Alur Standarisasi Entitas Nama

* + 1. Klasifikasi Jenis Kejadian

Setelah tweet informasi melalui proses ekstraksi informasi untuk mendapatkan entitas bernama beserta relasinya, maka proses ini adalah untuk menentukan jenis kejadian dari informasi tersebut. Klasifikasi jenis kejadian penting dalam penelitian ini dibagi dalam empat kelas yaitu; 1) kebakaran, 2) lalu-lintas (macet-Kecelakaan), 3) bencana alam (banjir, topan, gempa, dan longsor), 4) kejadian lainnya. Algoritma klasifikasi yang digunakan dalam klasifikasi ini adalah SVM *multi-class* dengan memanfaatkan fitur model *bi-gram* yang telah dijelaskan pada subbab sebelumnya.

* + 1. Visualisasi Data

Tahap terakhir dari sistem ini adalah proses visualisasi data. Tujuan dari visualisasi data adalah untuk memudahkan masyarakat dalam membaca informasi kejadian yang ada. Proses visualisasi data dibagi menjadi 2 tahap yaitu; 1) geocoding, dan 2) pengelompokan informasi kejadian yang sama. Pertama, Geocoding merupakan proses untuk mendapatkan koordinat lokasi kejadian (latitude dan lotitude) yang kemudian digunakan untuk memetakan lokasi kejadian dalam peta digital dengan menggunakan bantuan API Google MAPS sebagai tool visualisasi data. Kedua, pengelompokan data kejadian, merupakan proses pegelompokan informasi kejadian berdasarkan kesamaan lokasi, jenis, dan rentang waktu kejadian. Rentang waktu kejadian adalah selisih informasi awal dan informasi berikutnya yang merupakan satu informasi yang sama. Rentang waktu kejadian berupa satuan waktu yang ditetapkan disini sebesar 60 menit.

* 1. Ujicoba dan Analisa Hasil

Setelah tahapan desain sistem dibuat maka pada tahap ini dilakukan ujicoba sistem seperti pada Gambar 3.4 untuk mengetahui kemampuan metode deteksi kejadian pada data stream Twitter yang meliputi performa klasifikasi jenis tweet, klasifikasi jenis kejadian, dan ekstraksi informasi. Dalam penelitian ini, ada dua tahapan sistem yang perlu diuji yaitu tahap klasifikasi dan pengenalan entitas nama. Metode pengujian yang digunakan untuk menguji performa klasifikasi adalah perhitungan akurasi. Sedangkan metode uji yang digunakan untuk menguji performa pengenalan entitas nama adalah perhitungan *precision*, *recall*, dan *F-measure*.

* + 1. Akurasi

Akurasi digunakan untuk menghitung tingkat keakuratan sistem dalam melakukan klasfikasi data tweet kedalam kelasnya masing-masing. Rumus akurasi yang digunakan seperti pada persamaan 3.1.

(3.1)

* + 1. Precision, Recall, dan F-Measure

Precision adalah tingkat ketepatan antara entitas yang diminta untuk dikenali dengan jawaban yang diberikan oleh sistem. Sedangkan recall adalah tingkat keberhasilan sistem dalam mengenali entitas nama pada informasi.

Precision : perbandingan jumlah dokumen yang relevan terhadap query dengan jumlah dokumen yang terambil dari hasil pencarian. Rumus precision yang digunakan seperti pada persamaan 3.2.

(3.2)

Recall : perbandingan jumlah entitas relevan yang terambil sesuai dengan query yang diberikan dengan total kumpulan dokumen yang relevan dengan query. Rumus recall yang digunakan seperti pada persamaan 3.3.

(3.3)

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

(3.4)

* 1. Dokumentasi dan Jadwal Penelitian

Pada tahap dokumentasi akan dilakukan penulisan laporan hasil penelitian dari setiap tahapan yang dilakukan dalam penelitian. Tujuan dari tahapan ini adalah menghasilkan dokumentasi tertulis dari penelitian yang dilakukan secara sistematis. Jadwal penelitian yang dilakukan dapat dilihat pada.

Tabel . Jadwal Kegiatan Penelitian

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Kegiatan** | **Bulan** | | | | | | | | | | | | | | | | | | | |
| **Feb 18** | | | | **Maret 18** | | | | **April 18** | | | | **Mei 18** | | | | **Jun 18** | | | |
| 1. | Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. | Analisa dan Perancangan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. | Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. | Pengujian dan Evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Penyusunan Buku Tesis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# **DAFTAR PUSTAKA**

Kemendikbud, I. (2016) ‘Peraturan pendidikan dan kebudayaan Republik Indonesia’, *Kemendikbud*, (Standar Penilaian Pendidikan).

Lahitani, A. R. (2017) *Analisis Uji Metode Pembobotan Kata TF-IDF-DF dalam Upaya Meningkatkan Hasil Skoring pada Jawaban Esai Bahasa Indonesia*. Universitas Gadjah Mada.

Liang, G. *et al.* (2018) ‘Automated essay scoring: A siamese bidirectional LSTM neural network architecture’, *Symmetry*, 10(12), pp. 1–16. doi: 10.3390/sym10120682.

Tim Direktorat Pembinaan SMP (2017) ‘Panduan : Penilaian oleh Pendidik dan Satuan Pendidikan Sekolah Menengah Pertama’, p. 188.