

**SISTEM EKSTRAKSI INFORMASI OTOMATIS DARI *SUBTITLE*
VIDEO TUTORIAL *YOUTUBE***

SKRIPSI

Diajukan Guna Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S. Kom.)
Pada Program Studi Teknik Informatika



Oleh :

Moch Afaldo Danny Ardiansyah
NPM : 2213020159

**FAKULTAS TEKNIK DAN ILMU KOMPUTER
UNIVERSITAS NUSANTARA PGRI KEDIRI
2026**

HALAMAN PERSETUJUAN PEMBIMBING

Skripsi oleh:

Moch Afaldo Danny Ardiansyah
NPM : 2213020159

Judul :

SISTEM EKSTRAKSI INFORMASI OTOMATIS DARI *SUBTITLE* VIDEO TUTORIAL *YOUTUBE*

Telah Disetujui Untuk Diajukan Kepada Panitia Ujian/Sidang Skripsi
Program Studi Teknik Informatika Fakultas Teknik dan Ilmu Komputer
Universitas Nusantara PGRI Kediri

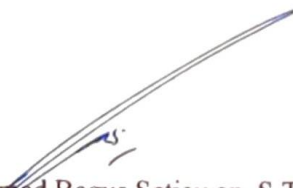
Tanggal : 14 Juli 2026

Pembimbing I



Ardi Sanjaya, M.Kom.
NIDN. 0706118101

Pembimbing II



Ahmad Bagus Setiawan, S.T., M.M., M.Kom.
NIDN. 0703018704

HALAMAN PENGESAHAN

Skripsi oleh:

Moch Afaldo Danny Ardiansyah
NPM : 2213020159

Judul :

SISTEM EKSTRAKSI INFORMASI OTOMATIS DARI *SUBTITLE* VIDEO TUTORIAL YOUTUBE

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi
Program Studi Teknik Informatika Fakultas Teknik dan Ilmu Komputer

Universitas Nusantara PGRI Kediri

Pada tanggal : 14 Juli 2026

Dan Dinyatakan telah Memenuhi Syarat

Panitia Penguji :

1. Ketua : Ardi Sanjaya, M.Kom.
2. Penguji I : Dr. Risky Aswi Ramadhani, S.Kom, M.Kom.
3. Penguji II : Ahmad Bagus Setiawan, S.T., M.M., M.Kom.



Dr. Sulistiono, M.Si
NIDN. 0007076801

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini saya,

Nama : Moch Afaldo Danny Ardiansyah
Jenis Kelamin : laki-Laki
Tempat/Tgl Lahir : 21-07-1999
NPM : 2213020159
Fakultas/Prodi : Teknik dan Ilmu Komputer/ Teknik Informatika

Menyatakan dengan sebenarnya, bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak dapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Kediri, 14 Juli 2026
Yang Menyatakan



The image shows a handwritten signature in black ink. To the left of the signature is a red circular stamp with the text 'KEMENTERIAN RISET, TEKNOLOGI, DAN INFORMATIKA' around the perimeter and 'KEMENTERIAN RISET, TEKNOLOGI, DAN INFORMATIKA' in the center. To the right of the signature is a red rectangular stamp with the text 'METERAI TEMPEL' and 'DE774AOX120292158' below it.

Moch Afaldo Danny Ardiansyah
NPM : 2213020159

HALAMAN PERSEMBAHAN

Penulisan skripsi ini dengan tulus saya dedikasikan kepada:

1. Kedua orang tua saya, yang dengan penuh kesabaran senantiasa mendoakan, memberikan dukungan terbaik, serta menjadi sumber motivasi tak henti-hentinya dalam perjalanan menyelesaikan skripsi ini.
2. Adik-adik saya, yang selalu memberikan semangat dan dukungan moral, sehingga saya dapat menyelesaikan proses ini dengan baik.
3. Seluruh dosen Universitas Nusantara PGRI Kediri, yang telah memberikan bimbingan, ilmu, dan pelajaran berharga, baik dalam ranah akademik maupun kehidupan sehari-hari.
4. Teman-teman seperjuangan di kampus, yang menjadi tempat berbagi suka dan duka, serta saling menyemangati selama menjalani masa perkuliahan hingga tahap akhir ini.
5. Almamater tercinta, Universitas Nusantara PGRI Kediri, yang telah menjadi tempat saya tumbuh, belajar, dan mengembangkan diri selama masa perkuliahan.
6. Semua pihak yang tidak dapat saya sebutkan satu per satu, yang telah memberikan kontribusi dalam berbagai bentuk dukungan selama proses penyusunan skripsi ini.

Semoga dedikasi ini dapat menjadi penghormatan atas semua dukungan, doa, dan kebersamaan yang telah diberikan

HALAMAN MOTTO

“When I look back at my life, I feel pain not because of suffering, but because of wasted time. I see how carelessly I lived, how often I ignored the quiet voice of my soul, how rarely I understood the value of a single moment. Only when death stood before me did I realize that life is not merely existence it is a miracle. Every minute is a treasure, and in every breath, there is the possibility of happiness.”

Fyodor Dostoevsky

“May you never lose the drive to pursue the things you love no matter how hard it gets” **Unknown**

“To love and lose and still to be kind” **Warsan Shire**

“What are you afraid of losing, when nothing in the world belongs to you.”

Marcus Aurelius

“Do not disturb yourself by imagining your whole life at once” **Marcus Aurelius**

RINGKASAN

Moch Afaldo Danny Ardiansyah Sistem Ekstraksi Informasi Otomatis dari Subtitle Video Tutorial YouTube, Skripsi, Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri, 2026

Kata Kunci : Ekstraksi Informasi, Subtitle YouTube, *Weak Supervision*, *Bootstrapping*, *Co-occurrence*, *Context Window*, *Triplet* Relasi, Bengkel Motor.

Penelitian ini bertujuan membangun sistem ekstraksi informasi otomatis dari subtitle video YouTube bertema perawatan dan perbaikan sepeda motor menggunakan kombinasi metode *Bootstrapping* dan *Weak Supervision*. Sistem dikembangkan untuk mengolah data transkrip lisan yang tidak terstruktur menjadi basis pengetahuan berupa relasi triplet terstruktur yang terdiri dari subjek, relasi, dan objek. Proses ekstraksi dilakukan melalui beberapa tahapan utama, yaitu pengunduhan subtitle otomatis menggunakan pustaka yt-dlp, pra-pemrosesan teks berupa pembersihan noise dan segmentasi kalimat, perluasan *seed list* menggunakan algoritma *bootstrapping* berbasis analisis *co-occurrence* terhadap 1.860 video korpus data, pelabelan lemah (*Weak Supervision*) dengan sistem pembobotan skor heuristik menggunakan ambang batas nilai 8, serta ekstraksi relasi menggunakan *context window* berukuran 15 kata untuk mendeteksi gejala dan tindakan di sekitar entitas komponen. Hasil ekstraksi divisualisasikan pada antarmuka *web* berbasis *Streamlit* yang dilengkapi fitur ekspor CSV. Hasil pengujian terhadap 16 video YouTube menghasilkan 71 data relasi triplet dengan nilai *True Positive* (TP) sebesar 66, *False Positive* (FP) sebesar 0, dan *False Negative* (FN) sebesar 5. Evaluasi menggunakan metrik *Precision*, *Recall*, dan *F1-Score* menghasilkan nilai masing-masing sebesar 100%, 92,66%, dan 96,35%, menunjukkan bahwa sistem mampu mengekstrak relasi informasi dari subtitle video YouTube bengkel motor dengan tingkat ketepatan dan keandalan yang sangat tinggi.

PRAKATA

Puji Syukur dipanjatkan kepada Tuhan Yang Maha Esa, karena atas ridha dan karunianya peneliti dapat menyelesaikan penyusunan laporan penelitian ini. Penulisan ini juga tak lepas dari dukungan pihak yang selalu membantu dalam penulisan penelitian ini. Oleh karenanya peneliti ingin mengucapkan terima kasih kepada:

1. Dr. Zainal Afandi, M.Pd., selaku Rektor Universitas Nusantara PGRI Kediri.
2. Dr. Sulistiono, M.Si., selaku Dekan Fakultas Teknik dan Ilmu Komputer Universitas Nusantara PGRI Kediri.
3. Ibu Risa Helilintar, M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Nusantara PGRI Kediri.
4. Ardi Sanjaya, M.Kom., dan Ahmad Bagus Setiawan, S.T., M.M, M.Kom. selaku Dosen Pembimbing Skripsi yang telah dan mengarahkan kami selama mengerjakan skripsi.
5. Kedua Orang Tua saya dan Keluarga atas doa dan dukungannya.
6. Ucapan terimakasih juga disampaikan kepada pihak-pihak lain yang tidak dapat disebutkan satu persatu, yang telah banyak membantu menyelesaikan penulisan penelitian ini.

Disadari penelitian ini masih banyak kekurangan, maka diharapkan kritik dan saran dari berbagai pihak sangat diharapkan. Semoga penelitian ini bermanfaat bagi semua pihak.

Kediri, 14 Juli 2026

Moch Afaldo Danny Ardiansyah
NPM : 2213020159

DAFTAR ISI

HALAMAN PERSETUJUAN PEMBIMBING	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN.....	iv
HALAMAN PERSEMBAHAN	v
HALAMAN MOTTO	vi
RINGKASAN	vii
PRAKATA	viii
DAFTAR ISI.....	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN.....	1
A. LATAR BELAKANG	1
B. Identifikasi Masalah.....	3
C. Rumusan Masalah.....	3
D. Batasan Masalah.....	3
E. Tujuan Penelitian	4
F. Manfaat dan Kegunaan Penelitian	5
BAB II LANDASAN TEORI	6
A. Teori dan Penelitian Terdahulu.....	6
1. Landasan Teori	6
2. Kajian Pustaka	11
B. Kerangka Berpikir	14
BAB III METODE PENELITIAN	16
A. Desain Penelitian.....	16
1. Jenis Penelitian (<i>Research Type</i>).....	16
2. Variabel Penelitian	17
3. Metode Pengumpulan Data	18

B. Instrumen Penelitian.....	19
C. Tempat dan Jadwal Penelitian.....	23
D. Objek Penelitian dan Subjek Penelitian	24
E. Prosedur Penelitian.....	27
1. Studi Literatur.....	27
2. Pengumpulan Data Sekunder	27
3. Pengumpulan Data Primer.....	28
4. Pengembangan dan Implementasi Sistem	28
5. Evaluasi	28
6. Analisis Hasil	28
7. Penyusunan Laporan	29
F. Teknik Analisis Data.....	29
1. Desain Sistem	29
2. Simulasi Proses Penyelesaian Masalah	39
BAB IV HASIL DAN PEMBAHASAN	50
A. Implementasi Sistem	50
1. Lingkungan Pengembangan.....	50
2. Pembangunan <i>Seed List</i> (Fase <i>Offline</i>)	50
3. Fase <i>Deploy</i> Sistem <i>Web</i>	55
4. Evaluasi Sistem	63
B. Pembahasan	64
BAB V PENUTUP	67
A. KESIMPULAN	67
B. SARAN.....	68
DAFTAR PUSTAKA	69
LAMPIRAN	71

DAFTAR TABEL

Tabel 3. 1 Rencana Kegiatan Penelitian	24
Tabel 3. 2 Contoh Data	40
Tabel 3. 3 Hasil Pra-pemrosesan.....	41
Tabel 4. 1 Perangkat Keras	50
Tabel 4. 2 Perangkat Lunak	50
Tabel 4. 3 Sumber dan Jumlah Data Video.....	51
Tabel 4. 4 Jumlah Seed Inisiasi Awal	52
Tabel 4. 5 Hasil Eksperimen Ukuran Sliding Window	53
Tabel 4. 6 Hasil Akhir Seed List.....	54
Tabel 4. 7 Input Link YouTube.....	55
Tabel 4. 8 Penanganan Error Input URL.....	56
Tabel 4. 9 Contoh hasil Weak Supervision.....	58
Tabel 4. 10 Tabel Triplet.....	59
Tabel 4. 11 Uji Coba Jumlah Context Window	59
Tabel 4. 12 Evaluasi Sistem	64

DAFTAR GAMBAR

Gambar 2. 1 Kerangka Berpikir	15
Gambar 3. 1 DFD Level 0.....	29
Gambar 3. 2 DFD Level 1.....	30
Gambar 3. 3 Diagram Dekomposisi.....	32
Gambar 3. 4 Flowchart.....	34
Gambar 3. 5 Class Diagram	36
Gambar 3. 6 Halaman Utama.....	38
Gambar 3. 7 Output Sistem.....	39
Gambar 4. 1 Tampilan Awal.....	61
Gambar 4. 2 Tampilan Ekstraksi.....	62
Gambar 4. 3 Tampilan Distribusi Komponen.....	62
Gambar 4. 4 Tampilan Ekspor Data.....	63

DAFTAR LAMPIRAN

Lampiran 1. 1 Kartu Bimbingan	73
Lampiran 1. 2 Surat Bebas Plagiasi	75
Lampiran 1. 3 Berita Acara Ujian Skirpsi.....	76
Lampiran 1. 4 Lembar Revisi.....	77

BAB I

PENDAHULUAN

A. LATAR BELAKANG

YouTube merupakan platform video dengan jumlah pengguna terbanyak di Indonesia, Dikutip dari laporan *We Are Social* (2025) pengguna youtube di Indonesia mencapai 143 juta orang. Selain sebagai media hiburan, Youtube juga menjadi rujukan utama masyarakat untuk mendapatkan sebuah tutorial termasuk dalam domain perbaikan motor. Kata kunci seperti “Penyebab motor *brevet*” secara konsisten selalu mendapatkan peringkat atas pencarian pada *Google Trends* sepanjang 2024-2025, namun sering kali video tutorial tersebut memiliki durasi panjang (10-30 menit) berisi penjelasan tidak langsung serta subtitle berbahasa indonesia informal, sehingga menyulitkan pengguna yang membutuhkan solusi cepat di lapangan.

Berdasarkan observasi lapangan di toko *spare part* sepeda motor (Mutiar Motor Nganjuk) Periode Agustus-Oktober 2025, karyawan yang tidak memiliki keahlian mekanik sering kali mendapatkan pertanyaan langsung dari konsumen mengenai gejala kerusakan kendaraan, Seperti “Motor Saya *Brevet* Kira-Kira bagian apa yang Rusak?” atau “Suara Mesin Motor Saya Kasar Biasanya Rusak Apa?”. Untuk menjawab dengan tepat sering kali karyawan toko membuka video tutorial di YouTube, Kemudian menonton video tersebut hingga selesai atau memindai *subtitle* secara manual. Proses tersebut memerlukan waktu cukup lama, menyebabkan antrean pelanggan memanjang serta mengurangi efisiensi pelayanan

Upaya untuk mengatasi permasalahan tersebut telah dilakukan pemanfaatan teknologi *Natural Language Processing* (NLP) untuk mengekstrak informasi dari teks yang tidak terstruktur. (Farhadi., 2013) sudah membuktikan bahwa *subtitle* YouTube dapat menjadi sumber data kaya untuk anotasi semantik melalui *Named Entity Recognition* (NER) dan *web-based NLP tools*. Sementara itu, menurut (Kusumawardani & Kusumawati., 2024) telah berhasil menerapkan NER berbasis *Bidirectional LSTM-CRF* pada teks medis

berbahasa Indonesia, dan (Darji et al., 2023) sudah menunjukkan keunggulan model *Transformer (BERT)* pada domain hukum.

Meskipun berbagai pendekatan NLP telah berhasil diterapkan, terdapat kesenjangan penelitian (*research gap*) yang cukup signifikan. Mayoritas penelitian terdahulu masih berfokus pada ekstraksi teks berbahasa Inggris yang relatif baku, atau teks berbahasa Indonesia formal pada domain terstruktur seperti medis dan hukum. Di sisi lain, belum ada penelitian yang secara khusus mengekstraksi informasi menggunakan teknik *Named Entity Recognition (NER)* dan *Relation Extraction (RE)* pada *subtitle* YouTube berbahasa Indonesia informal di domain perbengkelan sepeda motor. Karakteristik data pada domain ini memiliki tantangan tersendiri karena penuh dengan *slang* otomotif (bahasa bengkel), dialek lokal, singkatan, serta tingginya tingkat kesalahan transliterasi pada *subtitle* otomatis. Celah inilah yang menjadi rumusan masalah sekaligus kebaruan utama dalam penelitian ini..

Berdasarkan uraian permasalahan dan penelitian terdahulu yang telah dipaparkan, penelitian ini bertujuan untuk mengembangkan sistem ekstraksi informasi *otomatis* dari *subtitle* video tutorial *YouTube* berbahasa Indonesia pada domain perawatan dan perbaikan sepeda motor. Sistem yang diusulkan menerapkan teknik NER dan RE untuk mengidentifikasi entitas utama (gejala kerusakan, komponen, penyebab, dan solusi) serta hubungan di antaranya, sehingga menghasilkan ringkasan terstruktur berbentuk poin-poin singkat. Pengguna hanya perlu memasukkan tautan video *YouTube* untuk memperoleh intisari pengetahuan secara instan. Dengan demikian, karyawan toko *sparepart*, mekanik bengkel kecil, dan pemilik motor bisa mendapatkan jawaban yang dibutuhkan hanya dalam hitungan detik, tanpa harus menonton video panjang. Hal ini akan menghemat waktu, mempercepat pelayanan, dan mengurangi antrean pelanggan serta mendukung pengambilan keputusan di lapangan.

B. Identifikasi Masalah

Berdasarkan latar belakang tersebut, maka dapat diidentifikasi masalah sebagai berikut :

1. Pengguna (karyawan toko dan pemilik kendaraan) memerlukan waktu lama untuk memperoleh jawaban dari video tutorial otomotif di YouTube karena harus menonton video berdurasi 10 sampai 30 menit atau membaca subtitle yang panjang dan tidak langsung.
2. Belum ada sistem ekstraksi informasi berbasis NER dan RE yang secara khusus mampu menangani subtitle YouTube berbahasa Indonesia informal pada domain perbaikan sepeda motor, yang memiliki slang otomotif, singkatan, dan dialek lokal yang sangat berbeda dari teks formal atau subtitle bahasa Inggris yang sudah banyak diteliti.

C. Rumusan Masalah

Berdasarkan identifikasi masalah diatas, diusulkan rumusan masalah penelitian ini sebagai berikut :

1. Bagaimana merancang dan mengembangkan sistem ekstraksi informasi berbasis *Named Entity Recognition* (NER) dan *Relation Extraction* (RE) yang mampu memproses *subtitle* video YouTube berbahasa Indonesia informal pada domain perbaikan sepeda motor yang mengandung *slang* otomotif, singkatan, dan dialek lokal?
2. Bagaimana tingkat kemampuan sistem ekstraksi informasi berbasis NER dan RE tersebut mampu mengekstrak entitas utama (gejala kerusakan, komponen, penyebab, solusi) serta hubungan antar-entitas secara akurat sehingga dapat menghasilkan ringkasan terstruktur yang membantu pengguna memperoleh jawaban instan dari video tutorial YouTube berdurasi panjang?

D. Batasan Masalah

Berdasarkan permasalahan di atas, agar penelitian tetap terfokus dan tidak melebar, dilakukan pembatasan masalah sebagai berikut:

1. Data berupa *subtitle* video YouTube berbahasa Indonesia tentang perawatan dan perbaikan sepeda motor
2. Lokasi pengambilan data primer bersumber dari platform YouTube, sedangkan data sekunder diperoleh dari dokumen daring dan hasil observasi lapangan di toko *sparepart* sepeda motor Mutiara Motor Nganjuk
3. Entitas yang diekstrak terbatas pada gejala, komponen, penyebab, dan solusi.
4. Relasi yang diekstrak mencakup gejala, komponen, penyebab, dan solusi.
5. Bahasa Pemrograman yang digunakan adalah *Python* versi 3.10 atau di atasnya dengan *library* pendukung seperti *yt-dlp*, *NLTK*, *re*, *fuzzywuzzy*, dan *scikit-learn* serta *library* lain yang relevan untuk mendukung proses pengambilan data, prapemrosesan teks, dan ekstraksi informasi
6. Evaluasi kinerja sistem dilakukan pada level entitas dan relasi menggunakan metrik *Precision*, *Recall* dan *F1-Score* dengan membandingkan hasil ekstraksi sistem terhadap *ground truth* anotasi manual.
7. *Output* sistem berupa prototipe yang menerima input berupa tautan YouTube dan menghasilkan ringkasan terstruktur dalam bentuk poin-poin atau tabel.
8. Penelitian ini berfokus pada pengembangan dan evaluasi sistem ekstraksi informasi, sehingga tidak membahas aspek antarmuka pengguna secara mendalam maupun implementasi sistem dalam skala produksi

E. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, tujuan penelitian ini adalah:

1. Membuat sistem ekstraksi informasi berbasis *Named Entity Recognition* (NER) dan *Relation Extraction* (RE) yang mampu memproses *subtitle* video YouTube berbahasa Indonesia informal pada domain perawatan dan perbaikan sepeda motor yang mengandung slang otomotif, singkatan, dan campuran dialek lokal.
2. Mengimplementasikan sistem yang dapat mengekstrak entitas utama (Gejala, Komponen, Penyebab, Solusi) beserta relasi antar-entitas secara

otomatis dari *subtitle* video YouTube, sehingga dapat menghasilkan ringkasan terstruktur yang membantu mekanik, karyawan toko *sparepart*, dan pemilik kendaraan memperoleh jawaban instan tanpa harus menonton video berdurasi panjang.

F. Manfaat dan Kegunaan Penelitian

Manfaat Teoritis dan Praktis dari penelitian yang dilakukan adalah sebagai berikut:

1. Teoritis

Penelitian ini diharapkan dapat menambah referensi dalam bidang *Natural Language Processing* (NLP) untuk bahasa Indonesia informal, khususnya pada penerapan *Named Entity Recognition* (NER) dan *Relation Extraction* (RE). Selain itu, penelitian ini dapat menjadi dasar bagi penelitian selanjutnya yang ingin mengembangkan sistem ekstraksi informasi dari *subtitle* video.

2. Manfaat Praktis

Secara praktis, penelitian ini mempermudah mekanik, karyawan toko *sparepart*, dan pemilik kendaraan memperoleh informasi penting dari video YouTube tanpa harus menonton video penuh. Dengan adanya ringkasan informasi otomatis, penelitian ini diharapkan dapat meningkatkan efisiensi pelayanan dan mendukung pengambilan keputusan di lapangan.