

Penerapan Metode *Weak Supervision* dan *Open Information Extraction* untuk Ekstraksi Triplet Semantik pada *Subtitle* YouTube Otomotif

^{1*}Moch Afaldo Danny Ardiansyah, ²Ardi Sanjaya, ³Ahmad Bagus Setiawan

^{1,2,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: ¹aldodanny3@gmail.com, ²dersky@gmail.com, ³ahmadbagus@unpkediri.ac.id

Penulis Korespondens : Moch Afaldo Danny Ardiansyah

Abstrak— Pencarian informasi teknis perawatan otomotif melalui transkripsi otomatis YouTube sering terhambat oleh durasi video yang panjang dan derau bahasa lisan. Penelitian ini bertujuan mengembangkan sistem otomatis untuk mengekstrak pengetahuan diagnostik menggunakan kombinasi metode *Weak Supervision* dan *Open Information Extraction*. Data primer terdiri dari 1.790 dokumen subtitle YouTube, sedangkan data sekunder berupa kamus acuan istilah otomotif. Hasil pengujian terhadap 10 dokumen uji menunjukkan performa optimal dengan nilai *Precision* 100 persen, *Recall* 96,77 persen, dan *F1-Score* 98,36 persen. Sistem juga berhasil diintegrasikan ke dalam antarmuka web berbasis Streamlit untuk visualisasi data secara interaktif. Penelitian ini berkontribusi pada bidang pengolahan data berupa pengumpulan informasi efisien tanpa pemeriksaan manual. Celah informasi terlewat berupa dua *False Negative* dipicu oleh tidak terdeteksinya entitas suku cadang pada kamus acuan.

Kata Kunci—ekstraksi informasi, otomotif, *open information extraction*, *weak supervision*, youtube

Abstract—Searching for technical automotive maintenance information through automatic YouTube transcriptions is often hindered by long video durations and linguistic noise. This study aims to develop an automated system to extract diagnostic knowledge using a combination of *Weak Supervision* and *Open Information Extraction* methods. The primary data consists of 1,790 YouTube subtitle documents, while the secondary data is a reference lexicon of automotive terms. Test results on 10 documents demonstrate optimal performance with a *Precision* of 100 percent, *Recall* of 96.77 percent, and an *F1-Score* of 98.36 percent. The system is successfully integrated into a Streamlit-based web interface for interactive data visualization. This research contributes to the data processing field by providing an efficient information gathering method without manual inspection. Missing information in the form of two *False Negatives* was triggered by undetected spare part entities in the reference dictionary.

Keywords— write 3 to 5 words in alphabetical order and separated by commas

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Perkembangan platform digital telah mengubah lanskap pencarian informasi teknis secara mandiri oleh masyarakat. Proses pengambilan pengetahuan dari dokumen teks di internet saat ini menjadi tantangan besar karena mayoritas data digital hidup dalam bentuk teks tidak terstruktur.[1] Sebagai platform video dengan pengguna terbesar di Indonesia yang mencapai 143 juta orang berdasarkan laporan We Are Social tahun 2025, YouTube telah bergeser fungsi menjadi rujukan utama pembelajaran informal, termasuk dalam domain perawatan otomotif sepeda motor. Tren

pencarian pada Google Trends sepanjang tahun 2024–2025 secara konsisten menunjukkan tingginya kebutuhan publik terhadap topik pemecahan masalah kendaraan, salah satunya melalui kata kunci "*Penyebab motor brebet*". Namun, perolehan informasi secara instan dari platform ini sering kali terhambat. Video tutorial umumnya berdurasi panjang antara 10 hingga 30 menit, mengandung penjelasan yang tidak langsung pada inti masalah, serta didukung oleh transkripsi *subtitle* otomatis berbahasa Indonesia yang bersifat informal dan tidak terstruktur.[2], [3]

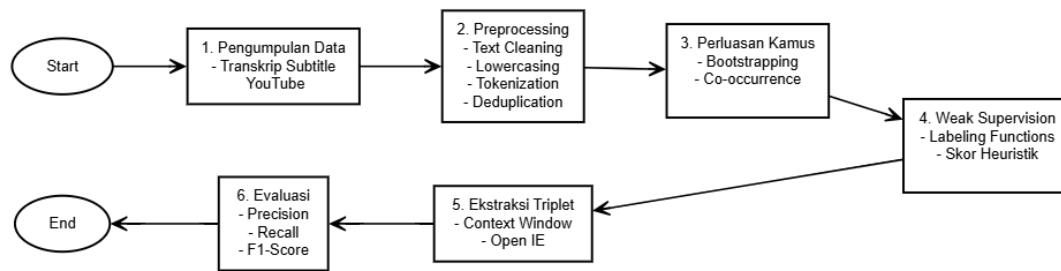
Kondisi tersebut berdampak langsung pada efisiensi operasional di lapangan. Studi kasus yang dilakukan di toko suku cadang Mutiara Motor Nganjuk menunjukkan bahwa karyawan yang tidak memiliki latar belakang mekanik sering kali kesulitan merespons pertanyaan diagnostik dari konsumen, seperti indikasi kerusakan komponen berdasarkan gejala suara mesin atau tarikan kendaraan yang tidak normal. Guna memberikan solusi yang tepat, karyawan terpaksa memindai teks *subtitle* atau menonton video tutorial YouTube secara manual di tempat. Proses penelusuran manual ini memakan waktu yang cukup lama, memicu penumpukan antrean pelanggan, serta menurunkan kualitas pelayanan toko. Oleh karena itu, diperlukan sebuah metode komputasi yang mampu mengekstrak intisari pengetahuan dari dokumen teks lisan tersebut secara cepat dan otomatis. Menurut Jurafsky dan Martin (2023), pendekatan *Information Extraction* (IE) di dalam ranah pemrosesan bahasa alami merupakan solusi krusial untuk mentransformasikan teks bebas yang tidak beraturan menjadi representasi data atau relasi formal yang terstruktur.[4]

Pemanfaatan teknologi *Natural Language Processing* (NLP) untuk ekstraksi informasi dari teks tidak terstruktur sebenarnya telah banyak dieksplorasi. Riset terdahulu membuktikan bahwa *subtitle* YouTube dapat dianalisis secara semantik menggunakan *Named Entity Recognition* (NER). Di Indonesia, implementasi NER berbasis deep learning seperti *Bidirectional LSTM-CRF* telah berhasil diterapkan pada dokumen konsultasi medis,[5] sementara model berbasis *Transformer* (BERT) menunjukkan performa unggul dalam domain hukum formal.[6] Meskipun demikian, mayoritas penelitian tersebut masih berfokus pada korpus berbahasa Inggris yang baku atau teks bahasa Indonesia formal. Hingga saat ini, belum ada model NER maupun *Relation Extraction* (RE) yang dikembangkan khusus untuk menangani *subtitle* YouTube berbahasa Indonesia informal pada domain perbengkelan sepeda motor. Korpus pada domain ini memiliki tantangan unik berupa tingginya penggunaan slang otomotif, singkatan teknis, interferensi dialek lokal, serta noise akibat kesalahan transkripsi otomatis (*Automatic Speech Recognition*).

II. METODE

Penelitian ini menggunakan pendekatan *Natural Language Processing* (NLP) untuk mengekstrak informasi diagnostik kendaraan dari *subtitle* video YouTube berbahasa Indonesia informal pada domain perbaikan sepeda motor. Metode yang digunakan adalah kombinasi *Weak Supervision* dan *Open Information Extraction* (*Open IE*) berbasis kalimat.

Weak Supervision bekerja dengan memanfaatkan fungsi pelabelan berupa aturan praktis berbasis *seed list* istilah otomotif untuk memberikan label awal secara otomatis pada baris kalimat. Sementara itu, *Open IE* bekerja dengan memproses struktur sintaksis kalimat tunggal untuk mengekstrak relasi semantik tanpa terikat pada ontologi kaku, sehingga menghasilkan luaran pengetahuan berbentuk triplet *Subject - Relation - Object*. Kombinasi metode ini dipilih karena memiliki kemampuan yang sangat baik dalam menangani karakteristik data teks lisan yang bersifat noisy, tidak terstruktur, penuh dengan slang otomotif, serta kesalahan transkripsi otomatis (*Automatic Speech Recognition*), seperti yang umum ditemukan pada korpus dokumen *subtitle* YouTube.



Gambar 2 Diagram Alir

1. Pengumpulan data

Proses pengumpulan data dalam penelitian ini dibagi menjadi dua kategori berdasarkan fungsinya yaitu data primer sebagai target ekstraksi informasi dan perluasan data sekunder sedangkan data sekunder sebagai basis pengetahuan awal.

Data Primer: Berupa korpus teks lisan sebanyak 1.790 dokumen subtitle video tutorial perbaikan sepeda motor yang dihimpun secara otomatis dari platform YouTube. Pemilihan data difokuskan pada konten kreator mekanik berbahasa Indonesia. Proses pengunduhan seluruh naskah subtitle dilakukan menggunakan pustaka *yt-dlp* dalam format *WebVTT*. Kumpulan dokumen mentah berskala besar inilah yang bertindak sebagai korpus target untuk diekstrak menjadi triplet relasi semantik dan selanjutnya digunakan sebagai perluasan Data Sekunder.

Tabel 1 Data Primer

Nama Saluran	Jumlah Subtitle
AGM Channel	265
APM Motor Chanel	300
kebumen motor1	519
mekanik roda dua chanel	706
Total Subtitle	1790

Data Sekunder: Berupa daftar istilah teknis otomotif roda dua yang dihimpun dari berbagai dokumen eksternal seperti katalog daftar suku cadang kendaraan serta catatan observasi istilah lokal yang lazim digunakan di bengkel konvensional lapangan. Data sekunder ini disaring dan dikelompokkan ke dalam tiga kategori kamus acuan utama, yaitu *Seed Sparepart*, *Seed Gejala*, serta *Seed Tindakan*. Kumpulan data acuan tersebut bertindak sebagai jangkar informasi dan landasan bagi fungsi pelabelan lemah pada arsitektur sistem.

Tabel 2 Data Sekunder

Nama Seed	Jumlah Kata
Seed Sparepart	38
Seed Tindakan	30
Seed Gejala	23

2. Preprocessing Data

Tahap prapemrosesan data bertujuan untuk mentransformasikan dokumen transkripsi mentah yang bersifat *noisy* menjadi unit-unit kalimat bersih. pada awalnya, sistem melakukan standarisasi terhadap seluruh berkas data primer yang masih berbentuk format *subtitle* dinamis untuk dikonversi menjadi dokumen teks murni. Menggunakan aturan ekspresi reguler, sistem menyaring dan mengeliminasi seluruh komponen non-leksikal penanda sistem YouTube yang tidak memuat konten informatif, seperti informasi *header* format *WEBVTT*, penanda durasi, tag pemformatan teks otomatis, nomor urut baris, hingga keterangan audio non-verbal seperti penanda teks musik atau tertawa.

Setelah dokumen bertransformasi menjadi berkas teks murni yang bersih dari atribut visual sistem, pemrosesan dilanjutkan pada penyeragaman karakter huruf menjadi huruf kecil secara menyeluruh. Teks panjang yang telah seragam tersebut kemudian disegmentasi menjadi kumpulan unit baris kalimat tunggal memanfaatkan karakter baris baru. Langkah segmentasi tingkat kalimat *sentence tokenization* ini penting untuk mengunci ruang lingkup analisis algoritma agar fokus mengekstrak relasi semantik hanya pada batasan struktur satu kalimat murni.

Setelah kalimat berhasil dipisahkan, sistem melakukan pemeriksaan kesamaan string secara sekuensial untuk mendeteksi dan mengeliminasi kalimat yang identik dengan kalimat yang sudah diproses sebelumnya. Langkah ini diterapkan untuk menghindari redundansi data akibat adanya kalimat berulang yang sering muncul pada transkripsi teks lisan informal.

Sebagai tahapan penutup dari prapemrosesan, sistem melakukan filtrasi linguistik dengan mengeliminasi kata kata stopwords standar bahasa Indonesia, yang diperluas dengan penghapusan kata filler words yang sering diucapkan dalam percakapan tutorial secara berulang (seperti kata '*mas bro*', '*sini*', '*nah*', '*tadi*', '*oke*').

3. Perluasan Kamus

Tahap pengayaan basis data kamus bertujuan untuk mengatasi keterbatasan cakupan kosakata awal (*seed list*) tanpa bergantung pada anotasi manual manusia yang memakan waktu. Proses ini berjalan secara otomatis di bagian *back end* dengan memanfaatkan korpus teks hasil prapemrosesan untuk menjaring kandidat entitas otomotif baru, yang meliputi komponen suku cadang, indikasi gejala kerusakan, serta tindakan penanganan mekanik. Metode yang diterapkan pada tahapan ini mengombinasikan teknik *bootstrapping* konteks dan analisis kemunculan bersama (*co-occurrence*) secara sekuensial yang diadopsi dari pengembangan sistem perluasan relasi berbasis *Snowball*. [7]

Pada langkah *bootstrapping* konteks, sistem memanfaatkan entitas suku cadang yang telah dikenali dalam kamus acuan awal sebagai kata jangkar. Sistem kemudian memindai kata-kata di sekitar jangkar tersebut menggunakan batasan radius jendela kata kontekstual (*context window*) berukuran lima kata di sekelilingnya. Kata-kata baru yang muncul di dalam jendela tersebut disaring secara ketat berdasarkan kriteria kelayakan, seperti mengeksklusi kata yang sudah terdaftar dalam kamus lama serta membatasi panjang karakter minimal tertentu agar terhindar dari kebisingan teks berupa singkatan yang tidak memiliki makna diagnostik. Kandidat kata yang lolos filter kemudian dihimpun secara otomatis berdasarkan kekuatan frekuensi kemunculannya.

Untuk memperkuat validasi dan mengukur tingkat asosiasi antar-kata secara objektif, kandidat istilah yang terjaring kemudian diuji lebih lanjut melalui analisis kemunculan

bersama (*co-occurrence*) dengan parameter *sliding window* berukuran sepuluh kata. Di dalam ruang lingkup jendela geser ini, hubungan kedekatan hubungan semantik antara istilah baru dengan kata pemicu dihitung secara matematis berdasarkan intensitas kemunculan bersama dalam satu konteks kalimat. Istilah baru yang berhasil melampaui ambang batas frekuensi kemunculan minimal sebanyak lima kali dinyatakan valid memiliki keterikatan domain yang kuat. Entitas baru yang telah tervalidasi tersebut kemudian diumpungkan kembali ke dalam repositori kamus acuan utama untuk memperbarui basis pengetahuan sistem.

4. *Weak Supervision*

Setelah repositori kalimat bersih berhasil dibangun pada tahap prapemrosesan, sistem tidak langsung melakukan ekstraksi relasi, melainkan melakukan penyaringan tingkat relevansi konteks menggunakan metode *weak supervision*. Tahapan ini diperlukan sebagai pelindung untuk menyingkirkan kalimat-kalimat umum, percakapan kasual, atau noise lisan yang tidak memuat konten informasi kegagalan mekanis, sehingga algoritma ekstraksi berikutnya dapat bekerja secara efisien pada korpus data yang tepat. Pendekatan programmatic berbasis aturan ini diadopsi untuk menggantikan pelabelan manual konvensional yang kerap menjadi hambatan terbesar (*bottleneck*) dalam siklus penyiapan data latih.[8]

Proses penyaringan ini bekerja dengan menerapkan fungsi pelabelan berbasis pembobotan skor heuristik.[9] Fungsi ini memanfaatkan kamus acuan kata jangkar yang telah diperluas pada tahap sebelumnya untuk memindai keberadaan entitas otomotif di setiap baris kalimat tunggal. Sistem memberikan bobot nilai tertinggi apabila mendeteksi adanya entitas komponen suku cadang (5), yang kemudian diikuti secara proporsional oleh bobot nilai untuk entitas gejala kerusakan (3) dan tindakan penanganan mekanik (2). Guna menangkap kalimat yang kaya akan struktur informasi, sistem menyuntikkan bonus skor (10) apabila entitas komponen terdeteksi muncul berpasangan langsung dengan entitas gejala atau tindakan dalam satu batasan kalimat yang sama. Kalimat dinyatakan valid dan diberi label relevan (*pseudo-label*) apabila akumulasi bobot nilainya berhasil memenuhi ambang batas kelayakan minimal (8). Hasil dari tahapan ini adalah sebuah kumpulan data kalimat relevan yang siap dibedah hubungan semantiknya

5. Ekstraksi Triplet *Open Information Extraction*

Kalimat-kalimat yang dinyatakan lolos kualifikasi dan telah berlabel relevan dari tahap *weak supervision* kemudian masuk ke dalam fase ekstraksi informasi utama. Pada tahap ini, sistem menerapkan prinsip *Open Information Extraction (Open IE)* tingkat kalimat yang dikombinasikan dengan metode *Context Window* untuk mentransformasikan frasa teks lisan menjadi struktur pengetahuan formal yang terstruktur. Pendekatan Open IE diadopsi karena kemampuannya dalam mengekstrak relasi semantik secara luas dari korpus teks tanpa bergantung pada skema ontologi yang kaku atau ditentukan sebelumnya.[10], [11]

Proses ekstraksi diawali dengan penerapan mekanisme *Context Window* untuk mengisolasi teks di sekitar entitas target. Filter ini bekerja dengan menetapkan parameter radius jendela kontekstual yang dinamis berukuran lima belas kata di sekeliling posisi koordinat entitas komponen suku cadang yang telah ditemukan. Melalui jendela isolasi ini, kata-kata di luar radius tersebut akan disembunyikan dari jangkauan pemindaian algoritma. Pendekatan ini sangat efektif untuk mengunci fokus pencarian gejala atau tindakan mekanik hanya pada untaian kata yang secara tekstual berada di dekat suku cadang terkait, sekaligus

meminimalkan ambiguitas akibat karakteristik bahasa lisan tutorial yang strukturnya sering kali melompat-lompat serta menghindari perambatan kesalahan (*error propagation*) akibat struktur sintaksis kalimat yang terlalu kompleks.[12]

Dalam arsitektur *Open Information Extraction* ini, penugasan peran semantik untuk setiap elemen pembentuk untaian pengetahuan didefinisikan ke dalam bentuk komponen relasi tuple atau triplet semantik [13]. Struktur elemen tersebut didefinisikan secara spesifik sebagai berikut :

Subjek : Berperan sebagai entitas jangkar informasi utama yang merepresentasikan nama komponen suku cadang sepeda motor yang berhasil teridentifikasi di dalam teks.

Relasi : Bertindak sebagai penghubung semantik yang menjembatani makna hubungan di dalam kalimat. Penentuan relasi ini dilakukan secara dinamis ke dalam dua kategori hubungan fungsional berdasarkan karakteristik kata pemicu (*trigger words*) yang terdeteksi di dalam jendela isolasi, yaitu relasi penunjuk indikasi kerusakan (mengalami gejala) ketika sistem menemukan kecocokan kata yang mengarah pada kondisi abnormal komponen, serta relasi penunjuk penanganan mekanis (dilakukan tindakan).

Objek : Berfungsi sebagai argumen penjelas deskriptif dari kondisi akhir yang menampung karakteristik kerusakan fisik, detail visual gejala abnormal, maupun tindakan penanganan konkret yang dialami atau dikenakan secara langsung pada subjek terkait[11].

6. Evaluasi

Kinerja sistem dievaluasi secara kuantitatif menggunakan parameter confusion matrix untuk menghitung metrik *precision*, *recall*, dan *F1-Score*. Rumus matematis yang digunakan untuk menghitung ketiga metrik evaluasi tersebut adalah sebagai berikut:

a. Precision

$$Precision = \frac{TP}{TP+FP} \dots \dots \dots (1)$$

b. Recall

$$Recall = \frac{TP}{TP+FN} \dots \dots \dots (2)$$

c. F1-Score

$$F1 = 2 \times \frac{Precision.Recall}{Precision+Recall} \dots \dots \dots (3)$$

Keterangan:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

III. HASIL DAN PEMBAHASAN

Pada bab ini disajikan seluruh luaran kuantitatif, contoh transformasi data dari setiap modul di dalam data pipeline, implementasi antarmuka sistem, hingga hasil pengujian performa menyeluruh serta analisis teoretis terhadap arsitektur sistem yang diusulkan.

3.1 Hasil Preprocessing

Tahap prapemrosesan data berhasil mentransformasikan naskah transkripsi otomatis platform YouTube yang bersifat acak dan sarat akan *noise* menjadi baris kalimat bersih yang siap diproses. Untuk menunjukkan perubahan struktur data secara transparan, diambil satu contoh baris data transkripsi mentah dokumen uji yang ditelusuri secara kontinu hingga tahapan akhir sebagai berikut:

Tabel 3 Hasil Preprocessing

Contoh Data		
	Data Kalimat Subtitle	Format File
Sebelum	kampas<00:02:26.6	.vtt
	90><c>	
	kopling</c><00:02	
	:27.049><c>	
	atau</c><00:02:27	
	.230><c>	
	kampas</c><00:02:	
	27.560><c>	
	ganda</c><00:02:2	
	7.860><c>	
Sesudah	habis</c>	.txt
	Kampas kopling atau kampas ganda habis	

3.2 Hasil Perluasan Data Sekunder

proses pengayaan kamus ini dipantau dan dikontrol secara ketat melalui fungsi filter panjang karakter serta ambang batas frekuensi kemunculan minimal >5. Langkah pemantauan ini krusial untuk memastikan tidak ada kebisingan teks (*noise*) yang lolos menjadi kosakata baru. total pertumbuhan baris data sekunder kamus acuan setelah proses perluasan disajikan pada Tabel 4 berikut :

Tabel 4 Data Sekunder Pasca Perluasan

Nama Seed	Jumlah Awal	Jumlah Pasca Perluasan
Seed Sparepart	38	61
Seed Tindakan	30	52
Seed Gejala	23	35

3.3 Hasil *Weak Supervision*

Tahap *weak supervision* berfungsi menyaring kalimat berkonten kegagalan mekanis dari repositori data bersih. Pada awal proses, sistem otomatis menggugurkan teks yang kurang dari 25 karakter guna mengeliminasi klausa pendek non-informatif. Kalimat yang lolos filter dimensi tersebut kemudian dinilai kelayakannya menggunakan fungsi pelabelan berbasis skor heuristik.

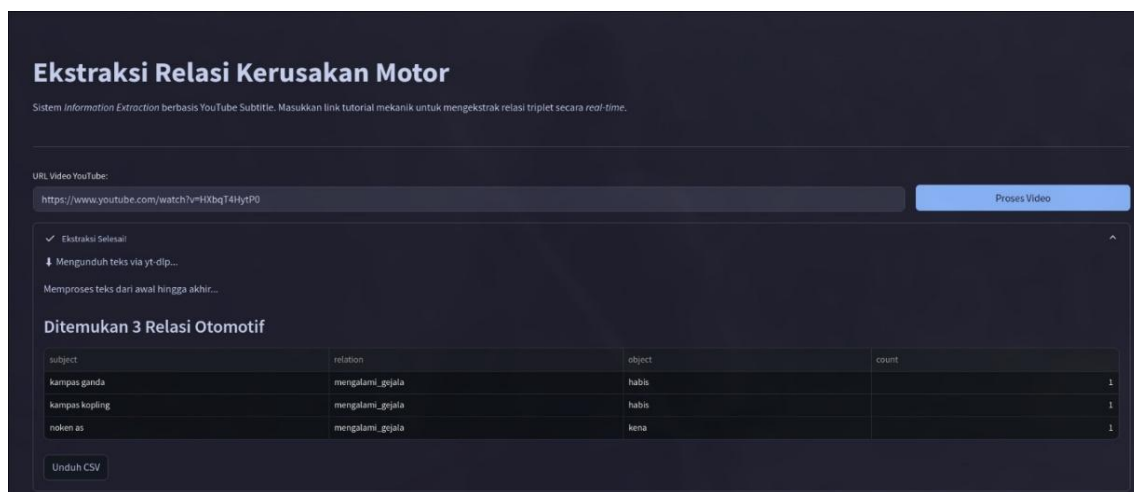
Sebagai contoh, hasil dari proses *weak supervision* pada salah satu data menunjukkan teks: "Kampas kopling kampas ganda habis.,2,1,0,23,"kampas ganda, kampas kopling"". Kalimat tersebut menghasilkan akumulasi skor sebesar 23. Sistem otomatis menetapkan status kalimat ini sebagai relevan (*pseudo-labeled*) karena telah melebihi nilai ambang batas yang ditetapkan.

3.4 Hasil Ekstraksi Triplet *Open IE*

Pada tahap ekstraksi informasi, kalimat yang lolos seleksi dibedah strukturnya menjadi bentuk pengetahuan formal. Sebagai contoh, hasil ekstraksi untuk contoh kalimat sebelumnya menghasilkan dua bentuk triplet semantik yaitu :

1. kampas ganda, mengalami_gejala, habis, 1
2. kampas kopling, mengalami_gejala, habis, 1

3.5 Implementasi Antarmuka Web



Gambar 2 Implementasi Sistem

3.6 Evaluasi Kinerja Sistem

Sistem diuji menggunakan 10 dokumen naskah tutorial perbaikan motor untuk mengukur akurasi ekstraksi triplet otomatis terhadap acuan manual (*Gold Standard*). Hasil pengujian parameter *confusion matrix* dan metrik performa disajikan pada Tabel 7 berikut.

Tabel 7 Evaluasi Sistem

Parameter Uji	Total Triplet	Metrik Evaluasi	Presentase
True Positive (TP)	60	Precision	100%
False Positive (FP)	0	Recall	96,77%
False Negative (FN)	2	F1-Score	98,36%

IV. KESIMPULAN

Penelitian ini berhasil membangun arsitektur *data pipeline* otomatis menggunakan integrasi metode *weak supervision* dan *open information extraction* (Open IE). Sistem ini mampu mentransformasikan transkrip subtitle video YouTube otomotif yang tidak terstruktur menjadi basis pengetahuan berbentuk triplet semantik (Subjek, Relasi, Objek) secara instan.

Berdasarkan pengujian menggunakan metrik standar *Natural Language Processing* (NLP), sistem menunjukkan performa komputasi yang sangat efektif dengan nilai presisi (*precision*) 100%, tingkat perolehan kembali (*recall*) 96,77%, dan akurasi harmonis (*F1-score*) 98,36%. Hasil ini diperoleh dari total 60 data *True Positive* (TP) dan hanya menyisakan 2 data *False Negative* (FN) tanpa adanya data salah tafsir atau *False Positive* (FP = 0). Bukti kuantitatif tersebut mengonfirmasi bahwa kombinasi skor heuristik fungsi pelabelan (*labeling functions*) dan isolasi *context window* mampu mereduksi derau lisan (*noise*) secara mutlak tanpa memerlukan anotasi manual manusia.

Meskipun sistem terbukti valid, munculnya 2 data *False Negative* (FN) menunjukkan bahwa performa ekstraksi akhir masih bergantung pada kelengkapan kamus acuan. Sebagai rekomendasi untuk penelitian selanjutnya, modul perluasan kamus secara iteratif perlu ditingkatkan dengan mengeksplorasi variasi kosakata lokal, singkatan mekanis, serta *slang* otomotif yang lebih luas. Hal ini krusial untuk meminimalkan risiko hilangnya data (*missing data*) akibat adanya entitas suku cadang baru yang belum terdaftar di dalam sistem.

DAFTAR PUSTAKA

- [1] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining." [Online]. Available: <http://www.crisp-dm.org/>
- [2] Ronen. Feldman and James. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.
- [3] B. Farhadi, "Enriching Subtitled YouTube Media Fragments via Utilization of the Web-Based Natural Language Processors and Efficient Semantic Video Annotations," 2013. [Online]. Available: <http://www.gjset.org>
- [4] D. Jurafsky and J. H. Martin, "Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft."
- [5] R. P. Kusumawardani and K. N. Kusumawati, "Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstmcrf algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1146–1156. doi: 10.1016/j.procs.2024.10.344.
- [6] H. Darji, J. Mitrović, and M. Granitzer, "German BERT Model for Legal Named Entity Recognition," Mar. 2023, doi: 10.5220/0011749400003393.
- [7] E. Agichtein and L. Gravano, "Snowball: Extracting Relations from Large Plain-Text Collections."
- [8] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment*,

Association for Computing Machinery, Nov. 2017, pp. 269–282. doi: 10.14778/3157794.3157797.

[9] S. Pawar, G. K. Palshikar, A. Jain, J. Bhat, and S. Johnson, “Weak Supervision using Linguistic Knowledge for Information Extraction,” NLPAL.

[10] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open Information Extraction from the Web.”

[11] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, “A Survey on Open Information Extraction,” Jun. 2018, [Online]. Available: <http://arxiv.org/abs/1806.05599>

[12] L. Cui, F. Wei, and M. Zhou, “Neural Open Information Extraction.” [Online]. Available: https://1drv.ms/u/s!ApPZx_

[13] S. Pawar, G. K. Palshikar, and P. Bhattacharyya, “Relation Extraction : A Survey,” Dec. 2017, [Online]. Available: <http://arxiv.org/abs/1712.05191>