

EKSTRAKSI INFORMASI DARI DOKUMEN CV MENGGUNAKAN PENDEKATAN HYBRID NATURAL LANGUAGE PROCESSING DAN RULEBASED

Oleh:

Syahfrizka Dyah Nazwa Umbara

Dosen Pembimbing: Ade Eviyanti

Informatika

Universitas Muhammadiyah Sidoarjo

Januari 2026



Pendahuluan

Latar Belakang:

Transformasi digital menuntut efisiensi rekrutmen. Volume pelamar yang masif membutuhkan *Applicant Tracking System* (ATS).

Masalah:

1. Variabilitas Format: Format CV sangat beragam (unstructured) dan sulit diproses.
2. Kelemahan Metode Lama:
 - *Rule-Based (Regex)*: Terlalu kaku, gagal jika format berubah sedikit saja.
 - *Deep Learning (BERT/LSTM)*: Membutuhkan resource komputasi (GPU) tinggi dan dataset besar.

Solusi: Pendekatan *Hybrid* (Jalan Tengah) untuk menyeimbangkan akurasi tinggi dan efisiensi komputasi.

Pertanyaan Penelitian (Rumusan Masalah)

Rumusan Masalah: Bagaimana membangun system ekstraksi CV yang akurat dan mampu menangani variasi format namun tetap ringan secara komputasi?

Tujuan Penelitian: Mengembangkan sistem ekstraksi informasi CV menggunakan metode Hybrid:

1. **Regex:** Untuk entitas faktual (Email, Telepon).
2. **NLP (Spacy):** Untuk entitas kontekstual (Skill, Experience)

Batasan Masalah

1. Sistem hanya berupa **prototipe**, bukan produk komersial.
2. Fokus pada **ekstraksi informasi dasar**: data pribadi, pendidikan, pengalaman, dan keterampilan
3. **Format CV** terbatas pada **PDF dan DOC/DOCX** berbahasa Inggris
4. Tidak mencakup **fitur pemeringkatan kandidat**.
5. **Evaluasi** difokuskan pada **akurasi ekstraksi**, bukan keamanan atau skalabilitas.

Ringkasan Penelitian Terdahulu

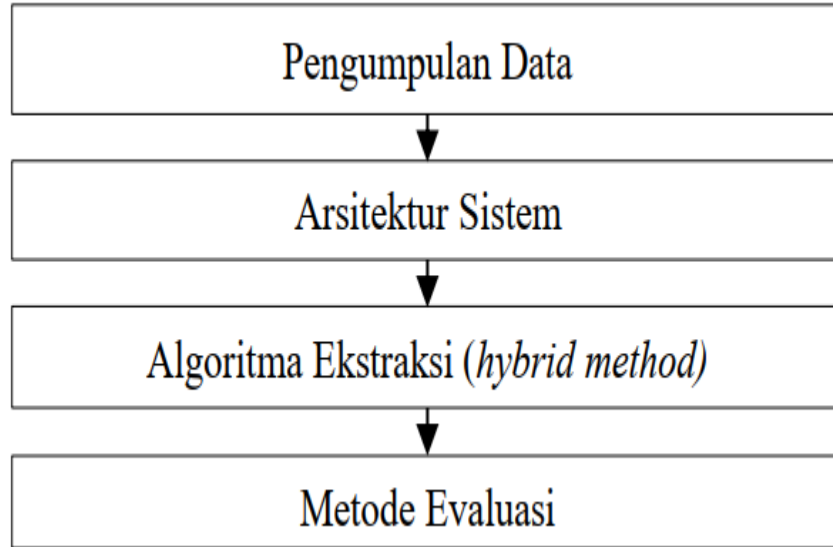
JUDUL PENELITIAN	METODE	KELEBIHAN	KEKURANGAN
Text Mining for Employee Candidates Automatic Profiling Based on Application Documents.	Text preprocessing, klasifikasi, feature extraction	Akurasi tinggi dalam klasifikasi profil	Hanya cocok untuk dokumen terstruktur
Application of NLP for Information Extraction from Unstructured Documents.	NER, rule-based extraction, text processing	Belum spesifik untuk CV	Efektif untuk dokumen tidak terstruktur
Analysis of Patterns and Similarities in Service Tickets using Natural Language Processing	Pattern analysis, similarity matching, text classification	Metode robust untuk analisis teks	Tidak fokus pada data personal
Analysis and shortcomings of e recruitment systems.	Semantic analysis, knowledge representation	Mengidentifikasi kekurangan sistem e-recruitment	Tidak memiliki implementasi teknis
AI-enabled interview Analysis	AI analysis, decision support system	Komprehensif untuk HRM	Fokus pada wawancara, bukan CV

Analisis Gap

1. **Evaluasi** akurasi **belum** dilakukan secara komprehensif dengan **metrik standar**.
2. **Belum menyesuaikan** dengan karakteristik khusus CV seperti **variasi format dan penulisan data**.
3. Umumnya menggunakan **satu metode** tunggal, belum mengintegrasikan berbagai **teknik NLP**.

Metode Penelitian

Alur penelitian



1. Pengumpulan Data:

Data Primer: Dokumen CV nyata (PDF/DOCX) yang dikumpulkan dari mahasiswa.

Data Sintetis: Dokumen CV buatan yang dirancang dengan variasi format khusus untuk menguji ketahanan sistem.

Data yang terkumpul kemudian dipetakan secara manual menjadi format JSON (*Ground Truth*) yang berfungsi sebagai kunci jawaban dalam proses evaluasi otomatis.

ADI PRAKOSO

Surabaya, East Java | +62 812 3456 7890 | adi.prakoso@email.com | [linkedin.com/in/adiprakoso](https://www.linkedin.com/in/adiprakoso)

SUMMARY

Final-year Informatics student at University of Airlangga with a specialized focus on Computer Vision and Deep Learning. A Bangkit Academy 2024 Graduate (Distinction) with experience in developing AI models for healthcare applications. Proven leadership as Head of Computer Science Student Association.

EDUCATION

- University of Airlangga | Bachelor's in Informatics | GPA: 3.88 | (2022–Present)
 - Activities: Head of Hima Informatika, Lab Assistant for Artificial Intelligence.

WORK EXPERIENCE

- Diskominfo Surabaya | Data Science Intern | (Aug 2025–Nov 2025)
 - Developed a dashboard to monitor city traffic patterns using Python and Tableau.
 - Cleaned and preprocessed large-scale municipal datasets for smart city analysis.
- PT Teknologi Cerdas Indonesia | AI Engineer Intern | (Feb 2025–July 2025)
 - Built a face recognition attendance system using OpenCV and TensorFlow with 95% accuracy.

SPECIALIZED TRAINING

- Bangkit Academy 2024 | Machine Learning Cohort
- Dicoding Indonesia | TensorFlow Developer Certification

SKILLS

Python, TensorFlow, Keras, OpenCV, SQL, Tableau, English (Proficient).

SYAHFRIZKA DYAH NAZWA UMBARA

Sidoarjo, East Java, 61215 | +62 895 8085 48030 | nazwaumbara@gmail.com
www.linkedin.com/in/nazwaumbara

Final-semester Informatics student at the University of Muhammadiyah Sidoarjo with proven leadership as Head of Pojok Statistik Agent at Umsida. A Bangkit Academy 2024 and Coding Camp 2025 graduate specializing in Machine Learning Engineering. Skilled in data analysis, Python programming, and Machine Learning model development with a focus on practical business solutions.

EDUCATION

University of Muhammadiyah Sidoarjo Sept 2022 – Present
Bachelor's Degree in Informatics | GPA: 3.94

- Lead of *Pojok Statistik Agent*, managing data-related initiatives to boost statistical literacy and student engagement in data-driven content.
- Contributed to Pojok Statistik Umsida's achievement of first place nationally in the BPS RI competition under the "Best BPS Regency-Guided Unit" category.
- Developed technical proficiency in Python, MySQL, MongoDB, and machine learning applications.

SMA Muhammadiyah 2 Jombang

Science Major 2019 – 2022

- Served as Class President and actively participated in student organizations to develop leadership and teamwork skills.
- Recognized as Best Graduate 2022 for academic excellence and organizational contribution.

WORK EXPERIENCE

Badan Pusat Statistik Sidoarjo Regency Oct 2025 - Present

Analysis Team – Survey Penyalahgunaan Narkoba Labatch 2025

- Prepared and documented comprehensive analytical reports to support evidence-based decision-making.
- Analyzed five core survey outputs to assess statistical trends and generate evidence-based insights.

Mitra Statistik 2025 Aug 2025 – Oct 2025

- Processed and visualized data maps using QGIS software.
- Performed data entry and verification to ensure accuracy and consistency.

Dinas Koperasi UKM Sidoarjo Regency

Intern – Secretariat & Cooperative Division May 2025 – Aug 2025

- Provided administrative support and assisted in data collection and report preparation.
- Supported cooperative evaluations and internal data management initiatives.

LAZISMU East Java Province

Freelance Web Developer (ZISKA project) Sept 2024 – Jul 2025

- Developed the frontend interface for the ZISKA web platform to improve user accessibility.

CV Wahana Karya Jaya

Intern – System Analyst Aug 2024 – Sept 2024

- Analyzed UI/UX, database structure, and backend functionality of the company's website.
- Provided recommendations for improving web performance and digital promotion strategies.

Freelance Graphic Designer

Feb 2024 – Aug 2025

- Created visual assets using Adobe Illustrator and Figma for various branding and marketing projects.

SPECIALIZED TRAINING

IDCamp 2025 Scholarship Powered by Indosat Ooredoo Hutchison

AI Multi-Platform App Developer Learning Path Sept 2025 – Present

- Developed foundational knowledge in artificial intelligence and cross-platform application development through hands-on coding and project-based learning.

Coding Camp Powered by DBS Foundation

Machine Learning Engineer Cohort Feb 2025 – Jul 2025

- Completed advanced ML training focusing on financial applications and real-world implementation.

Metode Penelitian

Pengumpulan Data



Arsitektur Sistem



Algoritma Ekstraksi (*hybrid method*)



Metode Evaluasi

2. Arsitektur Sistem:

Sistem dibangun berbasis web menggunakan kerangka kerja Flask (app.py). Alur pemrosesan data terdiri dari tiga modul utama:

1. Modul Input: Menerima file CV (PDF/DOCX) dan melakukan validasi format.
2. Modul Ekstraktor (extractor.py): Inti pemrosesan yang menggunakan library pdfplumber untuk konversi teks dan spaCy (en_core_web_md) untuk pemrosesan linguistik.
3. Modul Evaluator (evaluator.py): Mengukur akurasi hasil ekstraksi dibandingkan dengan data validasi manual.

3. Implementasi Algoritma:

1. Segmentasi Bagian (Logic)

Sebelum mengekstrak detail, sistem harus tahu mana bagian "Pendidikan" dan mana "Pengalaman".

```
self.headers_map = {  
    'education': ['education', 'academic', 'pendidikan', 'scholastic'],  
    'experience': ['work experience', 'employment', 'history', 'pengalaman'],  
    'skills': ['technical skills', 'skills', 'competencies', 'keahlian'],  
    'academy': ['academy', 'certifications', 'training', 'bootcamp', 'ach'],  
    'summary': ['professional summary', 'summary', 'profile', 'objective']  
}
```

2. Rule-Based (Regex)

Menggunakan Regular Expressions untuk data yang polanya pasti/baku.

```
def _extract_personal_info(self, text):  
    info = {}  
    email = re.search(r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b', text)  
    info['email'] = email.group(0) if email else None  
  
    phone = re.search(r'(\+62|62|08)(?:[s-]*\d){8,15}', text)  
    info['phone'] = phone.group(0).strip() if phone else None  
  
    li = re.search(r'(?:https?://)?(?:www\.)?linkedin\.com/in/[a-zA-Z0-9-]+', text, re.I)  
    info['linkedin'] = li.group(0) if li else None
```

Metode Penelitian

Pengumpulan Data

Arsitektur Sistem

Algoritma Ekstraksi (*hybrid method*)

Metode Evaluasi

3. NLP (PhraseMatcher)

Menggunakan spacy untuk mencocokkan *skill* teknis, termasuk kata majemuk (contoh: "Data Science").

```
self.skills_safe = [
    'Python', 'JavaScript', 'TypeScript', 'HTML', 'HTML5', 'CSS', 'CSS3',
    'SQL', 'MySQL', 'PostgreSQL', 'MongoDB', 'NoSQL', 'SQLite', 'Oracle', 'Redis',
    'Dart', 'Flutter', 'Android', 'iOS', 'Kotlin', 'Swift', 'Ruby', 'PHP', 'Laravel',
    'TensorFlow', 'PyTorch', 'Keras', 'Scikit-learn', 'Pandas', 'NumPy', 'Matplotlib',
    'Tableau', 'Power BI', 'QGIS', 'ArcGIS', 'Figma', 'Canva', 'Adobe Illustrator',
    'Git', 'GitHub', 'GitLab', 'Docker', 'Kubernetes', 'Jenkins', 'AWS', 'Azure',
    'Jira', 'Trello', 'Slack', 'Zoom', 'Google Workspace', 'Microsoft Office', 'Mac',
    'Data Science', 'Data Analytics', 'Computer Vision', 'NLP', 'Big Data', 'Statist
]

self.skills_strict = [
    'Go', 'React', 'Vue', 'Word', 'Excel', 'Java', 'C++', 'C#', 'R', 'C'
]

self.matcher_safe = PhraseMatcher(self.nlp.vocab, attr='LOWER')
self.matcher_strict = PhraseMatcher(self.nlp.vocab, attr='ORTH')

self.matcher_safe.add("SAFE", [self.nlp.make_doc(t) for t in self.skills_safe])
self.matcher_strict.add("STRICT", [self.nlp.make_doc(t) for t in self.skills_strict])
```

Evaluasi

dilakukan dengan membandingkan hasil output sistem (JSON) dengan file Ground Truth (ground_truth.json). Metrik yang digunakan dihitung secara otomatis oleh evaluator.py meliputi:

1. Precision: Tingkat ketepatan entitas yang diekstraksi.
2. Recall: Tingkat keberhasilan system menemukan seluruh entitas yang ada.
3. F1-Score: Rata-rata antara Precision dan Recall.

```
precision = tp / len(ext_set) if ext_set else 0
recall = tp / len(exp_set) if exp_set else 0
f1 = 2 * (precision * recall) / (precision + recall) if (precision + recall) else 0
return precision, recall, f1
```

KATEGORI ENTITAS	PRECISION	RECALL	F1-SCORE
Personal Info	1.00	1.00	1.00
Education	0.80	0.67	0.73
Skills	1.00	0.94	0.97
Experience	0.94	0.94	0.94
Academy	1.00	1.00	1.00

Implementasi Antarmuka Sistem

BATCH EXTRACTION

EVALUATION METRICS

Upload Multiple CVs

Pilih File

2 file

Analyze Batch

Select up to 10 PDF/DOCX files at once.

2 files selected

Halaman Utama

BATCH EXTRACTION

EVALUATION METRICS

System Performance Evaluation

Compare extraction results against Ground Truth JSON.

START EVALUATION

Precision

91.3%

Recall

87.5%

F1-Score

89.4%

Accuracy

80.8%

Filename	Precision	Recall	F1-Score	Accuracy
CV - Eng ver (8-10-2025).pdf	91.3%	87.5%	89.4%	80.8%

Halaman Evaluasi

CV - Eng ver (7-10-2025).pdf

SYAHFRIZKA DYAH NAZWA UMBARA

nazwaumbara@gmail.com +62 895 8085 48030 www.linkedin.com/in/nazwaumbara

Professional Summary

Final-semester Informatics student at the University of Muhammadiyah Sidoarjo with proven leadership as Head of Pojok Statistik Agent at Umsida. A Bangkit Academy 2024 and Coding Camp 2025 graduate specializing in Machine Learning Engineering. Skilled in data analysis, Python programming, and Machine Learning model development with a focus on practical business solutions.

Work Experience

Unknown Role

Badan Pusat Statistik Sidoarjo Regency

Aug 2025 – Oct 2025

Unknown Role

Dinas Koperasi UKM Sidoarjo Regency

May 2025 – Aug 2025

Intern – Secretariat & Cooperative Division

LAZISMU East Java Province

Sept 2024 – Jul 2025

Freelance Web Developer (ZISKA project)

CV Wahana Karya Jaya

Aug 2024 – Sept 2024

Freelance Graphic Designer

CV Wahana Karya Jaya

Feb 2024 – Aug 2025

Education

University of Muhammadiyah Sidoarjo Sept 2022 – Present

Bachelor's Degree in Informatics | GPA: 3.94

| GPA: 3.94

SMA Muhammadiyah 2 Jombang 2019 – 2022

Science Major

Technical Skills

Adobe Illustrator

Canva

Dart

Data Analytics

Deep Learning

Figma

Flutter

Google Workspace

Java

Machine Learning

Microsoft Office

MongoDB

MySQL

Python

QGIS

Statistical Analysis

Academy & Certifications

IDCamp 2025 Scholarship Powered by Indosat Ooredoo Hutchison

Participant

Coding Camp Powered by DBS Foundation Feb 2025 – Jul 2025

Participant

Bangkit Academy by Google, GoTo, Tokopedia, Traveloka

Sept 2024 – Feb 2025

Participant

CV 1 of 2

Visualisasi Hasil Ekstraksi

Kesimpulan dan Saran

Kesimpulan:

1. Integrasi metode Hybrid terbukti efektif menangani kompleksitas CV dengan sumber daya komputasi rendah.
2. Sistem mencapai akurasi 80.8% secara keseluruhan.

Saran Pengembangan:

1. Menambahkan fitur *Ranking Kandidat* otomatis berdasarkan kecocokan Job Description.
2. Memperbaiki logika ekstraksi Pendidikan untuk menangani format yang lebih unik.

Referensi

- [1] M. Saatci, R. Kaya, and R. Ünlü, "Resume Screening With Natural Language Processing (NLP)," *Alphanumeric J.*, vol. 12, no. 2, pp. 121–140, 2024, doi: 10.17093/alphanumeric.1536577.
- [2] K. Chhabra and D. Vashistha, "Applicant Tracking System (ATS) Department of Computer Science & Engineering and Information Technology Jaypee University of Information Technology ," vol. 173234, no. 201293, 2024.
- [3] Harshitha R and M. Veena, "A SURVEY ON RESUME ANALYSIS USING NLP," *www.irjmets.com @International Res. J. Mod. Eng.*, 1030, [Online]. Available: www.irjmets.com
- [4] N. Nair, S. Pavithra, and V. Vismaya, "Resume parser using NLP," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 13, no. 9, pp. 39–42, 2024, doi: 10.17148/IJARCCCE.2024.13905.
- [5] Y. Sari, M. F. Hassan, and N. Zamin, "Rule-based pattern extractor and Named Entity Recognition: A hybrid approach," *Proc. 2010 Int. Symp. Inf. Technol. - Eng. Technol. ITSIM'10*, vol. 2, pp. 563–568, 2010, doi: 10.1109/ITSIM.2010.5561392.
- [6] B. Nirali*, J. Gandhi, and D. K. Singh, "NLP based Extraction of Relevant Resume using Machine Learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 7, pp. 13–17, 2020, doi: 10.35940/ijitee.f4078.059720.
- [7] M. B. Gunjal, T. P. Thorat, K. S. Muttha, V. C. Shete, and P. D. Sagar, "A Review Paper on Resume Parser Using AI," 2025.
- [8] "Analysis & Shortcomings of E-Recruitment Systems: Towards a Semantics-based Approach Addressing Knowledge Incompleteness and Limited Domain Coverage 1."
- [9] S. Pudasaini, S. Shakya, S. Lamichhane, S. Adhikari, A. Tamang, and S. Adhikari, "Application of NLP for Information Extraction from Unstructured Documents," *Lect. Notes Networks Syst.*, vol. 209, pp. 695–704, 2022, doi: 10.1007/978-981-16-2126-0_54.
- [10] D. T. Tolciu, C. Săcărea, and C. Matei, "Analysis of patterns and similarities in service tickets using natural language processing," *J. Commun. Softw. Syst.*, vol. 17, no. 1, pp. 29–35, Feb. 2021, doi: 10.24138/JCOMSS.V17I1.1024.
- [11] R. Panwar, "AI-ENABLED INTERVIEW ANALYSIS: UNVEILING INSIGHTS AND ENHANCING DECISION-MAKING IN HUMAN RESOURCE MANAGEMENT," *INTERANTIONAL J. Sci. Res. Eng. Manag.*, vol. 07, no. 05, Jun. 2023, doi: 10.55041/IJSREM24357.
- [12] A. D. Wibawa, A. M. Amri, A. Mas, and S. Iman, "Text Mining for Employee Candidates Automatic Profiling Based on Application Documents," *Emit. Int. J. Eng. Technol.*, pp. 47–62, Apr. 2022, doi: 10.24003/emitter.v10i1.679.
- [13] R. Sonbol, G. Rebdawi, and N. Ghneim, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. The Use of NLP-Based Text Representation Techniques to Support Requirement Engineering Tasks: A Systematic Mapping Review", doi: 10.1109/ACCESS.2017.DOI.
- [14] E. Omodei et al., "OPEN ACCESS EDITED BY REVIEWED BY Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP." [Online]. Available: <https://thedeep.io>
- [15] M. F. Ghozali and A. Eviyanti, "Sistem Pakar Diagnosa Dini Penyakit Leukimia Dengan Metode 'Certainty Factor,'" *Kinetik*, vol. 1, no. 3, p. 135, 2016, doi: 10.22219/kinetik.v1i3.122.

