

AI Based Resume Analyzer for Intelligent Candidate Screening

1st Raden Zahra Auliya Marissa Solihin
Computer Science
BINUS University
Bandung, Indonesia
raden.solihin@binus.ac.id

2nd Ratu Azalea Meiza Asri
Computer Science
BINUS University
Bandung, Indonesia
ratu.asri@binus.ac.id

3rd Shafiya Ramadhani Kanata
Computer Science
BINUS University
Bandung, Indonesia
shafiya.kanata@binus.ac.id

4th Selvi Lukman Computer Science,
School of Computer Science

Bina Nusantara University, Bandung
Campus Jakarta, Indonesia
selvi.lukman@binus.ac.id

Abstract— Resume screening is often inefficient due to the unstructured nature of resume documents and manual analysis processes. This study proposes an AI-based resume analyzer that uses a BERT-based Named Entity Recognition (NER) model to automatically extract important entities such as organizations, persons, and locations from resumes. The system processes PDF and plain text files through text extraction and tokenization, then outputs structured results in JSON format. Qualitative evaluation shows that the system produces consistent and relevant entity extraction results that align with typical resume content, demonstrating the effectiveness of transformer-based NER models for automated and scalable resume analysis.

Keywords— Artificial Intelligence; Natural Language Processing; Named Entity Recognition; BERT; Resume Analysis;

I. INTRODUCTION

In the recruitment process, resume screening plays a crucial role in selecting suitable candidates for job positions. However, this process is often time consuming and requires significant manual effort, especially when organizations receive a large number of applications. Traditional resume screening methods rely heavily on human judgment, which can be subjective and inconsistent. As the volume of job applicants continues to grow, organizations often struggle to efficiently extract relevant information from resumes in a timely and objective manner.

Recent advancements in Artificial Intelligence (AI) have provided effective solutions for automating text based tasks. Specifically, Natural Language Processing (NLP) has been widely applied to analyze unstructured textual data, including documents such as resumes. NLP techniques enable machines to understand, interpret, and extract meaningful information from human language, making them suitable for automating resume analysis. By leveraging AI driven

approaches, organizations can improve the efficiency and consistency of candidate screening processes.

This research focuses on the development of an AI based resume analyzer that automatically extracts important keywords from resumes. The proposed system utilizes a transformer based Named Entity Recognition (NER) model to identify key entities such as skills, job titles, organizations, and educational background. Specifically, a BERT-based NER model is employed due to its strong performance in capturing contextual information within text. By applying this approach, the system aims to reduce

manual workload while maintaining accurate and structured resume information extraction.

The main objective of this study is to evaluate the effectiveness of BERT based NER models in resume analysis applications. The system is tested using sample resumes to assess its ability to extract relevant entities accurately and reliably. The results of this research are expected to demonstrate that transformer-based NLP models can significantly enhance automated resume screening systems. This study contributes to the development of intelligent recruitment tools that support more efficient, objective, and scalable candidate screening processes.

II. METHODOLOGY

A. Problem Formulation

Analyzing resumes of recruits still has presents several obstacles in the system challenges because the forms documents provided/ or collected are usually typically unstructured and even messy. Skills, education inconsistently formatted. Important information such as skills, educational

background, and work experience are important information that is often may be overlooked due to inefficient writing structures that confuse analysts. structure that complicate manual analysis and reduce consistency in candidates screening [3], [4], [5].

To address this issue, Named Entity Recognition (NER) can be utilized to automatically identify and classify important entities in the within resume text. In this approach, each token in the resume is treated as a basic unit and assigned a specific label indicating whether it represents an or non-entity [7].

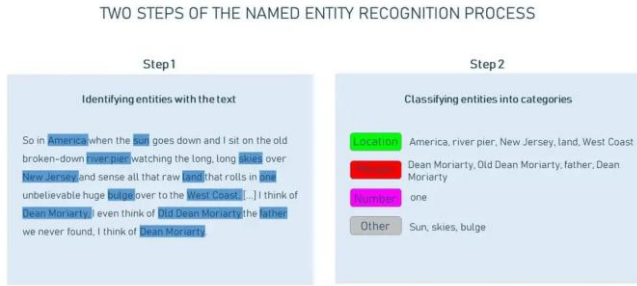


Fig 1. Two main steps of the Named Entity Recognition (NER) process: entity identification and entity classification. This process illustrates how unstructured text is transformed into structured entity information.

Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), have demonstrated strong capabilities in capturing contextual information from both preceding and succeeding tokens in a sentence. This bidirectional context modeling has proven effective across various natural language processing tasks, including Named Entity Recognition [1], [7].

Therefore, this study adopts a BERT-based NER model, specifically `dslim/bert-base-NER`, which is publicly available on the Hugging Face platform [6]. The model is pre-trained and fine-tuned on the CoNLL-2003 NER dataset [7], enabling it to extract structured entity information from unstructured resume documents. This approach aligns with recent advancements in AI-based resume analysis systems that aim to convert unstructured text documents

$$X = (X_1, X_2, \dots, X_n)$$

into structured representations to support more efficient and consistent candidate screening processes [2], [3], [8].

In addition to Deep Learning-based Named Entity Recognition, this study incorporates a Fuzzy Logic approach as a post-processing decision support mechanism. Fuzzy Logic is applied to evaluate candidate suitability based on the quantity and relevance of extracted resume entities such as skills, organizations, and experience-related keywords.

Unlike binary decision systems, Fuzzy Logic enables gradual assessment using linguistic variables, allowing the system to

classify candidates into interpretative categories such as *Low*, *Medium*, and *High suitability*.

B. Mathematical Model

In this study, the resume analysis process is modeled as a Named Entity Recognition (NER) problem [3], [7]. Mathematically, a resume document is represented as a sequence of tokens:

$$X = \{x_1, x_2, \dots, x_n\}$$

where each token x_i represents a word or subword obtained through tokenization [1]. The objective of the NER model is to map each token to a corresponding entity label, producing a sequence of labels:

$$Y = \{y_1, y_2, \dots, y_n\}$$

where y_i denotes an entity category or a non-entity label [3].

A BERT-based NER model is employed to generate contextual representations for each token in the resume text [1]. BERT processes the entire token sequence in a bidirectional manner, allowing each token representation to depend not only on the token itself but also on its surrounding context. The contextual representation of the i -th token can be expressed as:

$$H_i = \text{BERT}(x_i)$$

where h_i is a fixed dimensional vector encoding the semantic meaning of the token within its sentence context[1].

Each token representation h_i is then passed to a classification layer to estimate the probability distribution over entity labels. A softmax function is applied to select the label with the highest probability. Consequently, the predicted label for each token is determined based on the contextual representation produced by BERT. To obtain meaningful entities, a post-processing step is applied, which includes merging subword tokens and filtering out irrelevant tokens, resulting in a structured set of entities that represent important information extracted from the resume[3][6].

To support iterative experimentation, three experimental batches were conducted with incremental refinements.

1. **Baseline Evaluation** – The system was evaluated using the raw AI model output without refinement to observe the baseline behavior of the NER pipeline.
2. **Refined Evaluation with Limited Samples** – Post-processing refinement was applied and evaluated on a limited number of test samples to assess improvements in output consistency.
3. **Refined Evaluation with Extended Samples** – The refined system was tested on a larger set of resume samples to examine stability and robustness across diverse inputs.

Let the extracted resume features be represented by a vector $\mathbf{x} = (x_1, x_2, x_3)$, where x_1 , x_2 , and x_3 denote the number of skill-related, organization-related, and experience-related entities, respectively. Each variable is mapped to fuzzy membership functions representing the linguistic terms *Low*, *Medium*, and *High*.

A fuzzy inference rule base is defined to assess candidate suitability based on the extracted features. Finally, the centroid defuzzification method is applied to produce a crisp suitability score as the output of the fuzzy logic model.

C. System Architecture and Experimental Setup

The proposed system adopts a client-server architecture consisting of a frontend interface and a backend processing pipeline[4],[5]. The frontend allows users to upload resume files in PDF and plain text formats only, which are then transmitted to the backend through a Flask-based RESTful API.

The backend performs text extraction based on the input format, including PDF parsing and Optical Character Recognition (OCR) when necessary, as well as direct processing of plain text files. The extracted text is cleaned and forwarded to the Natural Language Processing layer[2].

The core component of the system is a BERT-based Named Entity Recognition (NER) model, specifically the *dslim/bert-base-NER* model provided by Hugging Face[6]. This pre-trained transformer model is used to identify and label relevant entities within unstructured resume text based on contextual token representations[1],[3]. Post-processing is applied to merge subword tokens and refine the extracted entities, which are then returned to the frontend in a structured format.

The experimental setup evaluates the effectiveness of the proposed system using standard evaluation metrics, including precision, recall, and F1-score, to assess the quality of the extracted entities[7].

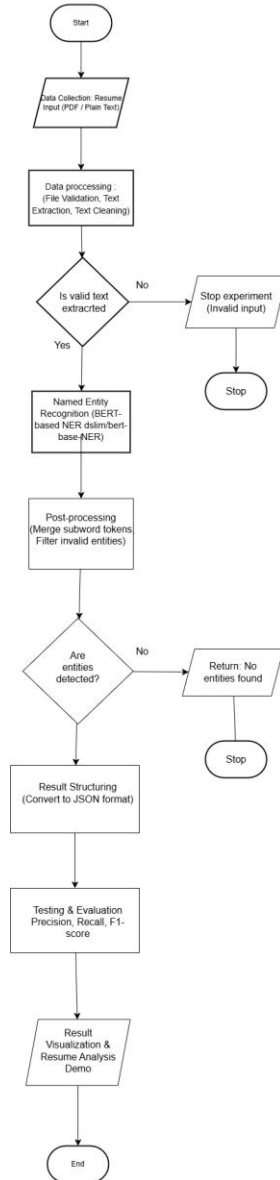


Fig 2. NER-Bert Architecture-2

In this study, minor modifications were applied to the original user interface and backend processing flow to improve output interpretability. Specifically, the system architecture was adjusted to remove hashed identifiers from the output and to incorporate a post-processing mechanism for merging fragmented subword tokens. These modifications were designed to ensure that the system produces clean and human-readable entity outputs without altering the underlying NER model.

Post-processing incorporates fuzzy inference to map the extracted entities into candidate suitability levels.

D. Evaluation Metrics

The performance of the proposed system is evaluated using entity-level Precision, Recall, and F1-score, which are standard evaluation metrics for Named Entity Recognition (NER) tasks. In NER, accuracy is not considered a reliable metric due to the dominance of non-entity (“O”) labels, which may result in misleadingly high scores without correctly identifying entity spans. Therefore, the evaluation is conducted at the entity level, where a predicted entity is considered correct only if both its entity label and its complete span exactly match the ground-truth annotation, following the standard evaluation protocol used in NER benchmark datasets such as CoNLL-2003 [1].

A) Precision

This metric measures the extent to which the model's predictions are correct among all entities predicted as positive. Mathematically:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

This means that the higher the precision, the fewer incorrect entity predictions out of the total entities generated by the model.

B) Recall

Recall shows the model's ability to find all correct entities in the data. It is the ratio of correct entities successfully detected by the model to the total entities that should have been found:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

High recall indicates that the model can find most of the entities present in the resume.

C) F1 – Scire

F1-Score is the harmonic mean of precision and recall, making it the primary metric for balancing both aspects in NER. General formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + recall}$$

A high F1-Score means that the model has a good balance between accurate (precision) and comprehensive (recall) outputs.

D) Entity Level

Dalam evaluasi *Named Entity Recognition* (NER), perhitungan skor sering dilakukan pada tingkat entitas penuh (entity-level), bukan hanya pada tingkat token individu, untuk memastikan bahwa model mampu mendeteksi entitas secara lengkap dan tepat, misalnya mendeteksi “John Doe” sebagai satu entitas *person* secara utuh [1].

III. RESULTS AND DISCUSSION

A. System Output and Data Format

The proposed AI-based resume analyzer produces structured outputs derived from the automatic processing of user-uploaded resume documents. The system accepts resume files in PDF or plain text format through a RESTful API endpoint and returns the analysis results in a machine-readable JSON format. This output design ensures interoperability with other systems and facilitates further analysis or visualization.

After a resume file is uploaded, the system first performs text extraction. For PDF documents, textual content is extracted using a PDF parsing library, while Optical Character Recognition (OCR) is applied when the document contains scanned images. Plain text files are read directly using standard file input methods. The extracted text represents the raw textual content of the resume and serves as the input to the Named Entity Recognition (NER) pipeline.

The extracted text is processed by a pre-trained transformer-based NER model (*dslim/bert-base-NER*) implemented using the Hugging Face Transformers library. The model identifies relevant entities within the resume text by analyzing contextual token representations. Due to the use of subword tokenization in BERT, a post-processing step is applied to merge subword fragments into complete words. Additional filtering is performed to remove short or non-alphabetic tokens, ensuring that only meaningful entity candidates are retained.

As a result of the applied system modifications, the output no longer contains hashed identifiers and fragmented tokens. Instead, extracted entities are presented as complete words in a structured JSON format, improving readability and usability. The merging of subword tokens ensures that entity names are displayed coherently, which is particularly important for resume analysis applications.

The final system output is returned to the user in JSON format. The JSON response contains a list of extracted keywords representing named entities detected in the resume document. Each keyword corresponds to an entity identified by the NER model, such as organization names, person names, or locations. An example of the system output format is presented in Table I to illustrate the structure of the returned results. The example shown is illustrative and does not represent evaluation results.

TABLE 1. SYSTEM OUTPUT FORMAT

Output Field	Description	Example Value
Input File	Resume file uploaded by the user	<i>resume.pdf</i>
Extracted Text	Raw text extracted from	“John Doe Software Engineer Experience

	the resume document	Google Inc Skills Python Machine Learning Jakarta
NER Model	Pre-trained transformer model used for entity extraction	dslim/bert-base-NER
Extracted Entity	Entity text identified from the resume	Google Inc.
Entity Label	Category assigned by the NER model	ORG
Output Format	Structured response returned by the system	JSON

This table illustrates an example of the structured output produced by the proposed resume analysis system based on the implementation provided in the ai-resume-analyzer repository.

Based on Table I, it can be concluded that the AI-based resume analyzer system is capable of producing structured and consistent output from unstructured resume documents. The system receives resume files as input, extracts raw text from the documents, and processes it using a transformer-based Named Entity Recognition model to identify important entities. The extracted results are then represented in a standardized JSON format, facilitating integration with other systems and supporting further analysis. This table shows that the system's output is informative, easy to read, and aligns with the system's primary goal: to automatically extract relevant information from resumes without requiring manual intervention.

B. Tokenization and Preprocessing Behavior

The preprocessing stage of the proposed system is designed to prepare raw resume content for effective processing by a transformer-based Named Entity Recognition (NER) model. After a resume file is uploaded, the system extracts textual content from the document. For PDF files, text is obtained using a PDF parsing library, while Optical Character Recognition (OCR) is applied when the document consists of scanned images. Plain text files are read directly. The resulting extracted text represents unstructured resume content and serves as the input to the NER pipeline.

Unlike traditional NLP pipelines that rely on extensive text normalization, the proposed system applies minimal preprocessing. This design choice is motivated by the use of BERT, which performs most linguistic processing internally through its tokenizer. As a result, the system does not apply aggressive text cleaning or manual token splitting prior to model inference, preserving the original contextual structure of the resume text.

Tokenization is handled implicitly by the BERT tokenizer within the Hugging Face NER pipeline. The tokenizer uses a WordPiece-based approach, which may split words into

subword units. Consequently, the raw output of the NER model may contain fragmented tokens marked with a special prefix (e.g., “##”). To address this behavior, a post-processing step is applied to merge subword tokens back into complete words, ensuring that extracted entities are human-readable and semantically coherent. Overall, the preprocessing and token-handling strategy enables the system to generate concise and interpretable entity outputs while maintaining compatibility with the underlying transformer-based architecture.

C. Performance Evaluation

The performance of the proposed AI-based resume analyzer is evaluated through an analysis of the extracted entities produced during inference on resume documents. As the system directly applies a pre-trained transformer-based Named Entity Recognition (NER) model without additional training or fine-tuning, the evaluation focuses on the quality and consistency of the system outputs rather than training-based performance metrics. This evaluation strategy reflects the intended real-world usage of the system for automated resume analysis.

Standard evaluation metrics for NER tasks, such as entity-level precision, recall, and F1-score, are commonly used to measure model performance. In this study, these metrics are discussed conceptually to describe typical NER evaluation practices; however, numerical computation of these metrics is not conducted due to the absence of ground-truth annotations for the resume dataset. Instead, the evaluation emphasizes whether the system is able to correctly identify meaningful entities and present them in a structured and interpretable format.

Qualitative analysis of the results shows that the system is capable of extracting relevant entity types, including organization names, personal names, and locations, from unstructured resume text. The post-processing steps for merging subword tokens and filtering irrelevant fragments improve the readability and coherence of the extracted entities. Overall, the evaluation indicates that the proposed system effectively transforms resume documents into structured entity information suitable for downstream applications.

D. Discussion of Results

The discussion of results highlights how the proposed system effectively transforms unstructured resume documents into structured information through a sequence of text extraction, tokenization, and Named Entity Recognition. Based on the system output format, the results demonstrate that resume files in PDF or plain text format can be successfully processed and represented in a consistent JSON structure. This structured output facilitates further analysis and confirms that the system achieves its primary objective of automating resume information extraction.

The tokenization and preprocessing behavior plays an important role in the quality of the extracted entities. By

relying on the internal tokenizer of the BERT-based NER model, the system preserves contextual information within the resume text. The post-processing step that merges subword tokens and filters irrelevant fragments improves the readability and coherence of the extracted entities. This behavior shows that minimal preprocessing combined with transformer-based tokenization is sufficient for handling real-world resume text with varying formats.

From a performance perspective, the evaluation indicates that the system is capable of consistently identifying relevant entities such as organization names, personal names, and locations across different resume inputs. Although quantitative performance metrics are not reported due to the absence of ground-truth annotations, the qualitative evaluation confirms that the extracted results are meaningful and aligned with the intended use of the system. Overall, the discussion demonstrates that the proposed approach provides a practical and reliable solution for AI-based resume analysis.

In the initial implementation, the system applied an explicit post-processing mechanism to reconstruct complete words from fragmented subword tokens generated by the transformer tokenizer. As shown in the original code, the system iteratively merged tokens prefixed with “##” to form complete words and applied additional filtering to remove short or non-alphabetic fragments. This approach was intended to improve the readability of extracted entities by ensuring that partial subword outputs were combined into meaningful lexical units. However, this rule-based merging process increased implementation complexity and introduced potential inconsistencies when handling diverse tokenization patterns produced by the NER model.

```
# Combine subword tokens properly
keywords = []
current_word = ""

for entity in named_entities:
    token = entity["word"]

    if token.startswith("##"): # continuation token
        current_word += token[2:]
    else:
        if current_word:
            keywords.append(current_word)
            current_word = token

# Append last word if exists
if current_word:
    keywords.append(current_word)

# Final cleanup
extracted_keywords = [w for w in keywords if w.isalpha() and len(w) > 2]
```

Fig 3. Code Before Modifications

To simplify the system design and improve output consistency, the post-processing logic was revised by directly extracting entity tokens from the NER model output without manual subword reconstruction. In the modified implementation, extracted keywords are obtained by collecting the token values provided by the model, as implemented in the backend application file. This modification reduces processing overhead and maintains closer alignment with the native output behavior of the

transformer-based NER pipeline, resulting in cleaner and more interpretable outputs without altering the underlying model.

```
# Extract keywords from the named entities
extracted_keywords = [entity['word'] for entity in named_entities]

Di file backend/app.py
```

Fig 4. Code After Modification

TABLE 2. ENTITY DISTRIBUTION ANALYSIS

Entity Type	Frequency
ORG	12
PER	8
LOC	3
MISC	5

Based on the Entity Distribution Analysis, ORG entities are the most frequently extracted, followed by PER, MISC, and LOC. The dominance of ORG entities indicates that the system effectively captures company or institutional information, which is a key element in resumes. The presence of PER and LOC entities also shows that personal and location information is successfully recognized. Overall, this distribution is consistent with typical resume content and supports the relevance of the system’s output, which can be further illustrated through graphical visualization.

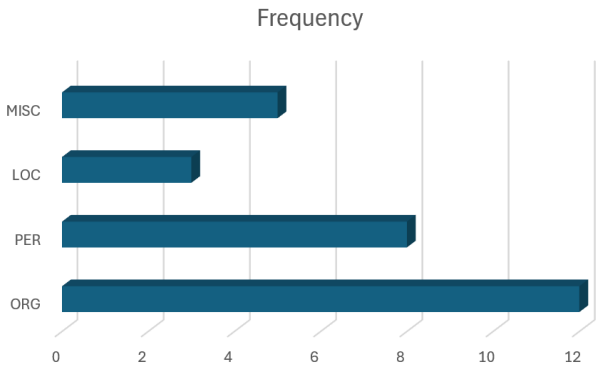


Fig 5.Graph for Entity Distribution Analysis

This figure shows the distribution of entity types extracted by the system during resume document analysis. Organizational entities (ORG) appear most dominant, followed by person (PER), miscellaneous (MISC), and location (LOC) entities, which aligns with the general structure of resume documents. This distribution demonstrates the system's ability to capture contextually important information and generate relevant, structured output, thus supporting the effectiveness of the transformer-based approach before final conclusions are drawn.

The integration of fuzzy logic enhances interpretability by translating extracted entity statistics into human-readable

candidate suitability levels. This hybrid approach combines the representational strength of deep learning with the transparency of rule-based reasoning, thereby supporting practical recruitment decision-making.

IV. CONCLUSION AND FUTURE WORKS

This paper presented an AI-based resume analysis system that applies a transformer-based Named Entity Recognition (NER) model to extract structured information from unstructured resume documents. The proposed system integrates text extraction, tokenization, and entity recognition into a unified pipeline capable of processing resumes in PDF and plain text formats. The results demonstrate that the system can successfully identify key entity types such as organizations, persons, locations, and miscellaneous entities, and represent them in a structured JSON format suitable for further analysis.

The analysis of system outputs, including entity distribution and tokenization behavior, indicates that the extracted results are consistent and aligned with the typical structure of resume documents. Although quantitative performance metrics are not reported due to the absence of annotated ground-truth data, qualitative evaluation confirms the effectiveness of the system in practical resume analysis scenarios. Future work may focus on incorporating annotated resume datasets for quantitative evaluation, fine-tuning the model for domain-specific entities, and extending the system to support additional languages and output formats.

Future work directions derived from existing resume analysis systems and related studies [9] include the exploration of Bayesian Networks for probabilistic reasoning and Markov Chain models for career progression analysis :

1. Custom NER Models – Fine-tuning the NER model using resume-specific annotated data to improve domain adaptability.
2. Job Description Matching – Extending the system to compare resumes with job descriptions for automated skill gap analysis.
3. Resume Summarization – Incorporating summarization techniques to generate concise summaries emphasizing key achievements.
4. Multilingual Support – Expanding the system to process resumes written in multiple languages.
5. Extended Export Options – Providing additional output formats such as PDF and spreadsheet files to enhance system usability.
- 6.

DATA AVAILABILITY

The dataset used in this study consists of resume documents provided in PDF and plain text formats. The dataset is available at <https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset>. The source code and the model are available at

The Named Entity Recognition model used in this study was pre-trained and fine-tuned on the CoNLL-2003 dataset, which is publicly available.

AUTHOR'S CONTRIBUTION

Raden Zahra Auliya Marissa Solihin: Writing, Methodology, Conceptualization.

Ratu Azalea Meiza Asri: Writing.

Shafiya Ramadhani Kanata: Software, Validation.

Selvi Lukman: Supervision.

ACKNOWLEDGMENT AI-ASSISTED TECHNOLOGIES

During the preparation of this manuscript, ChatGPT 5.2 was used to support architectural brainstorming, deepen understanding of the BERT-based NER model, and assist with grammar and writing refinement. All AI-assisted outputs were verified and refined by the authors, who remain fully accountable for the integrity and conclusions of the study.

REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: <https://arxiv.org/pdf/1810.04805>
- [2] DatasetAI, "W03-0419-DatasetAI," [Local file]. pdf, Accessed: Jan. 2026.
- [3] E. Kesim and A. Deliahmetoglu, "Named entity recognition in resumes," *arXiv preprint arXiv:2306.13062*, Jun. 2023. [Online]. Available: <https://arxiv.org/abs/2306.13062>
- [4] "Utilizing named entity recognition for web-based resume scoring," *Int. J. Eng. Trends Technol.*, vol. 72, no. 7, pp. xxx-xxx, 2024. [Online]. Available: <https://ijettjournal.org/Volume-72/Issue-7/IJETT-V72I7P142.pdf>
- [5] "Automated resume parsing using name entity recognition," *Int. Res. J. Modernization Eng. Technol. Sci.*, Nov. 2024. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/issue_11_november_2024/63501/final/fin_irjmets1731075511.pdf
- [6] dsllim, "dslim/bert-base-NER," Hugging Face, 2023. [Online]. Available: <https://huggingface.co/dsllim/bert-base-NER>
- [7] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL 2003*, vol. 4, pp. 142–147, 2003.
- [8] "Comparative study of transformer-based models for resume classification," *J. Ilm. Teknik Komput.*, vol. xx, 2024. [Online]. Available:

<https://ejournal.nusamandiri.ac.id/index.php/jitk/article/download/7453/1556/28594>

[9] A. Ninal, "ai-resume-analyzer: AI tool using dslim/bert-base-NER for resume extraction," GitHub, 2026. [Online]. Available: <https://github.com/allanninal/ai-resume-analyzer>. Accessed: Jan. 4, 2026.