Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.10 No.1 January 2025, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Entity Extraction and Annotation for Job Title and Job Descriptions Using Bert-Based Model

Anindo Saka Fitri[1], Seftin Fitri Ana Wati[2], Herlambang Haryo Putra[3], Suryo Widodo[4], Arizia Aulia Aziiza[5]

[1,2]*Information System Department, Universitas Pembangunan Veteran Jawa Timur, Indonesia*

[3]*Arion Digital Media, Indonesia*

[4]*Mathematics Education Department, Universitas Nusantara PGRI Kediri, Indonesia*

[2]*Information System Department, Universitas Surabaya, Indonesia*

[2]seftin.fitri.si@upnjatim.ac.id (*)

[1]anindo.saka.si@upnjatim.ac.id, [3]herlmabangharyo@gmail.com, [4]suryowidodo@upnkediri.co.id, [5]ariziaaulia@staff.ubaya.ac.id

Received: 2023-11-23; Accepted: 2025-01-20; Published: 2025-01-31

Abstract— This research paper investigates Named Entity Recognition (NER) within Indonesia's job vacancy domain, employing state-of-the-art Bert-based models. The study presents a detailed data collection and preprocessing methodology, followed by the Bert-based model's fine-tuning for enhanced NER. The dataset comprises 48,673 job vacancies collected from the JobStreet website in July 2023, specifically focusing on multi-entity recognition, including job titles and job descriptions. An original annotation algorithm was developed using Python and Laravel for precise entity recognition. In addition, this paper provides an extensive literature review of NER and Bert-based models and discusses their relevance in the context of the Indonesian job market. The outcomes highlight the efficacy of our BERT-based model, attaining an average accuracy of 78.5%, a precision of 79.7%, a recall of 81.1%, and an F1 score of 80.8% in the Named Entity Recognition (NER) task. The study concludes by discussing the implications, limitations, and future directions, underscoring the model's potential applicability in streamlining job matching and recruitment processes in Indonesia and beyond. This research contributes to the field by providing a robust framework for NER in job vacancies, highlighting the potential for improved job matching, and proposing enhancements for future model development and application in other languages and regions.

Keywords— Named Entity Recognition; Bert-Based Model; Job Vacancy; Deep Learning; Language Processing.

I. INTRODUCTION

The ability to automatically extract meaningful information from textual data is a fundamental aspect of natural language processing (NLP) with wide-ranging implications across various industries and applications. In the era of digitalization and information abundance, the volume of unstructured text data has grown exponentially [1]. Extracting structured knowledge from this vast sea of text is a formidable challenge but one that offers immense value. By enhancing efficiency, improving accuracy, and reducing human errors, structured data extraction saves resources and limits mistakes [2]. Named Entity Recognition (NER) is an essential task in natural language processing that focuses on detecting and classifying specific entities in text, including individuals, organizations, places, dates, and other similar elements.

In job vacancies, NER plays a pivotal role in efficiently matching job seekers with relevant job opportunities, streamlining the recruitment process, and enhancing user experience [3][4]. The job market is a dynamic and rapidly evolving landscape characterized by many entities crucial for job matching. Job titles, educational requirements, company names, and job descriptions are just a few examples of named entities within job postings that demand accurate identification and categorization.

However, NER for job vacancies presents unique challenges, especially when dealing with the Indonesian language, characterized by diverse linguistic features and a wide variety of named entities. The Indonesian language is known for its morphological richness, flexibility, and informal language variations, making it a complex environment for NERClick. Tap here to enter text. Job descriptions often contain specific domain-related terms and jargon, further complicating the NER task.

This research aims to address the specific challenge of NER for job vacancies in Indonesia by leveraging Bert-based models, which have demonstrated exceptional performance in NER tasks across various languages [5][6][7][8][9]. We seek to enhance the accuracy and efficiency of named entity recognition within job vacancies by applying these state-of-the-art NLP models to the Indonesian context. Our study focuses on a comprehensive dataset of 48,673 job vacancies collected from the JobStreet website in July 2023, covering many entity types crucial for the job market.

II. RESEARCH METHODOLOGY

This section delves into the existing body of literature that informs the research on Named Entity Recognition (NER) and its application in the context of job vacancies, particularly within the Indonesian language. The research flow in Fig.1 includes data collection and preprocessing, annotation, model training, and evaluation. This methodological framework is proposed to develop a robust NER system tailored to the Indonesian job market, aiming to enhance the efficiency of the region's job search and recruitment processes.

DOI : https://doi.org/10.25139/inform.v10i1.7367

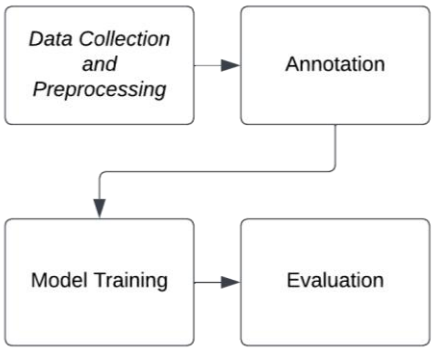

Fig.1. Research Methodology

A. Data Collection and Preprocessing

The data collection phase was carried out by scraping job vacancies from the JobStreet website in July 2023, resulting in a comprehensive dataset of 48,673 records. During this phase, we focused on capturing multi-entity relationships, specifically targeting job titles and job descriptions due to their interconnected nature with other entities [13].

Following the collection, the dataset underwent an extensive data-cleaning process to ensure the accuracy and consistency of the information. This included removing duplicates, correcting errors, and standardizing formats. Additionally, we generated a new dataset highlighting the most frequently occurring bi-grams and tri-grams within the job postings. This analysis helped identify common phrases and terms in job descriptions and titles. Furthermore, another dataset containing individual word segments was created, enabling a more granular analysis of the language used in job vacancies. These additional datasets provided valuable insights into the linguistic patterns and terminologies in the job market.

B. Annotation

The named entities within the dataset encompassed various aspects of job vacancy, including educational requirements, minimum and maximum age requirements, technical skills, personality traits, certifications, job prerequisites, job responsibilities, company names, job industries, job positions, technical skills, job placement (ranging from country to village levels), placement areas, buildings, employment statuses, minimum experience categories, job descriptions, and job application procedures. These entities were subsequently transformed into a new labelling scheme following the BIO (Beginning, Inside, Outside) format, enabling more effective NER model annotation and training.

The annotation process was executed using a custom-built algorithm developed in Python and Laravel, allowing for precise and context-aware entity recognition. Laravel was designed to provide an interface for understanding the context of job vacancies, which can be annotated from bi-grams and tri-grams. This interface also facilitated the labelling process. For example, a tri-gram found in the context of job postings, such as "*jurusan sistem informasi*" (information systems major), would be labelled as O-B-I. Python was then used to execute this labelling on the word segment data, ensuring the labelled dataset accurately recognized and tagged the named entities within the job vacancies. This tailored approach facilitated the creation of a labelled dataset that was essential for training and evaluating our model's performance.

C. Model Training

The model training phase in this research is a critical stage in developing the Named Entity Recognition (NER) system for job titles and job descriptions using Bert-based models. This phase was conducted with meticulous attention to data management. The data was divided into three subsets to ensure an effective training process. Approximately 80% of the data was allocated for training purposes, serving as the backbone for fine-tuning the Bert-based model to suit the specific requirements of NER within the context of Indonesian job postings.

In order to evaluate and improve the model's performance, the remaining 20% of the data was split into two subsets: 10 % was used for validation, and 10% was used for testing. The validation set was utilized during the training process to track the model's performance and fine-tune hyperparameters as needed. On the other hand, the test set was kept separate and used to evaluate the model's final performance and generalize to unseen data.

A detailed evaluation process was employed throughout the training phase to monitor the model's performance. To evaluate the model's performance in accurately identifying and classifying named entities in job postings, metrics like accuracy, precision, recall, and F1-score were employed. The model underwent training over multiple epochs, and checkpoints were saved at different stages to ensure the availability of the best-performing models.

The results of the model training phase hold significant implications for the overall success of the NER system. The trained model is anticipated to excel in recognizing entities within job titles and descriptions, aiding job seekers and employers in matching job requirements and qualifications. This phase sets the stage for the subsequent evaluation and testing of the NER system to ascertain its effectiveness in real-world scenarios.

1) Named Entity Recognition (NER): NER is a well-established field in natural language processing (NLP) with a wide range of applications, from information retrieval to question-answering systems and sentiment analysis [10] [11]. In the realm of NER, several studies have demonstrated the effectiveness of Bert-based models in capturing contextual information for entity recognition in various languages. Additionally, previous research has explored NER in the context of the job market. Still, the challenges posed by job vacancies, such as informal language, abbreviations, and unique entity types, have received limited attention. To the best of our knowledge, this study is one of the first to address the task of NER in Indonesian job vacancies by leveraging Bert-based models. It draws upon the foundation laid by previous NER research and adapts these techniques to suit the nuances of job vacancies in Indonesia. Our work builds upon the

DOI : https://doi.org/10.25139/inform.v10i1.7367

insights gained from the broader NER literature and tailors them to the specific domain of job advertisements in the Indonesian language, ultimately contributing to the advancement of NER techniques in a real-world, application-driven context.

2) Bert-Based Models: The Bert-based model, short for Bidirectional Encoder Representations from Transformers, is a state-of-the-art deep learning architecture developed by Google. It has gained prominence in the NLP community due to its ability to understand the contextual relationships between words in a sentence by considering both left and right context, as opposed to earlier models that only looked at one direction. This bidirectional understanding of language allows Bert-based models to excel in various NLP tasks, including NER. They have proven their adaptability across languages and domains, making them an attractive choice for our study.

Additionally, previous research [12] has explored NER in the context of the job market. Still, the specific challenges posed by job vacancies, such as informal language, abbreviations, and unique entity types, have received limited attention. To the best of our knowledge, this study is one of the first to address the task of NER in Indonesian job vacancies by leveraging Bert-based models. It draws upon the foundation laid by previous NER research and adapts these techniques to suit the nuances of job vacancies in Indonesia. Our work builds upon the insights gained from the broader NER literature and tailors them to the specific domain of job advertisements in the Indonesian language, ultimately contributing to the advancement of NER techniques in a real-world, application-driven context.

In the context of Named Entity Recognition (NER) and natural language processing (NLP), Bert-based models have emerged as a groundbreaking development. “Bert” stands for “Bidirectional Encoder Representations from Transformers,” and it represents a neural network architecture created by Google. Bert-based models have garnered substantial attention in the NLP community due to their remarkable ability to understand the contextual relationships between words in a sentence.

Unlike earlier models that relied on unidirectional language modelling, Bert-based models take a bidirectional approach. This means they consider both left and right contexts when processing words, resulting in a more comprehensive language understanding. This bidirectional capability enables them to capture nuances, context, and dependencies within a text, making them exceptionally effective for various NLP tasks, including NER.

One of the primary reasons for the success of Bert-based models is their pre-training on large corpora of text data. During pre-training, these models learn to predict missing words within sentences, effectively acquiring a profound understanding of language structures. This pre-trained knowledge is then fine-tuned for specific NLP tasks, such as NER, making Bert-based models adaptable and transferable to different languages and domains.

In this study, we employ Bert-based models as the core technology for NER in job vacancies, as they have proven their effectiveness in capturing contextual information and understanding the complexities of the Indonesian language. This section has provided an overview of the fundamental principles of Bert-based models and their relevance to our research, setting the stage for their application in Indonesia's specific domain of job advertisements.

D. Evaluation

To assess the effectiveness of our approach, we conducted rigorous evaluations of the model’s performance. We employed standard NER metrics such as precision, recall, and F1 score to quantify the model’s ability to recognize named entities accurately. We also conducted qualitative analyses to understand the model’s strengths and weaknesses, which provided insights into potential areas for improvement.

III. Result and Discussion

The Results section uses Bert-based models to present the key findings and outcomes of the Named Entity Recognition (NER) system for job titles and job descriptions. This section encompasses both quantitative and qualitative assessments of the model’s performance.

The model’s performance was rigorously evaluated using standard NER metrics, including accuracy, precision, recall, and F1-score. These metrics are vital for correctly assessing the system’s ability to recognize and categorize named entities within job postings. The impressive accuracy, precision, recall, and F1-score values demonstrate the model’s effectiveness in identifying named entities within job titles and descriptions. These results underscore the NER system’s potential to streamline matching job requirements and qualifications, benefiting job seekers and employers.

Table I
Model Performance

Metric	Value
Accuracy	78,5
Precision	79,7
Recall	81,1
F1-Score	80,8

The impressive accuracy, precision, recall, and F1-score values demonstrate the model’s effectiveness in identifying named entities within job titles and descriptions. These results underscore the NER system’s potential to streamline matching job requirements and qualifications, benefiting job seekers and employers.

In addition to quantitative metrics, qualitative assessments were performed by manually inspecting a representative subset of recognized entities—the qualitative analysis aimed to ensure the accuracy and relevance of the named entities identified by the model.

The combination of quantitative and qualitative evaluations provides a comprehensive understanding of the NER system’s performance, making it a valuable tool for enhancing job search and talent recruitment processes.

DOI : https://doi.org/10.25139/inform.v10i1.7367

The results presented in this section form the basis for the paper's discussion and conclusion, where the findings' significance is further explored and implications for future work are discussed.

The results of our research shed light on several significant insights and implications for the field of Named Entity Recognition (NER) in the context of Indonesian job vacancies and natural language processing (NLP) in general. In this section, we explore the significance of our results, acknowledge the constraints of our study, and examine possible directions for future research and advancements.

Our Bert-based NER model has demonstrated remarkable accuracy and adaptability in recognizing named entities within job vacancies, contributing to the advancement of NER techniques in the Indonesian job market. The accuracy of 78.5%, precision of 79.7%, recall of 81.1%, and F1 score of 80.8%, respectively, showcase the model's potential practical utility in enhancing job matching and recruitment processes.

However, it's crucial to acknowledge that our model's performance may vary across different sub-domains of job vacancies or specific industries, as it primarily relies on the data it was trained on. Future research may involve domain-specific fine-tuning or ensemble approaches to further enhance its effectiveness in diverse job markets. Moreover, the NER task is sensitive to the quality and diversity of the training data, and ongoing efforts to improve the dataset's quality should be considered for even more accurate entity recognition.

One of the notable limitations of our study is the relatively small dataset for individual named entities in standalone testing. This resulted in overlap between tasks and technical skills and between job placement entities. Furthermore, the entity "tasks" may warrant further investigation as it can disambiguate between other entities and hold potential for more in-depth analysis in future research.

While our study marks a significant step forward, it also raises important questions about the generalizability of Bert-based models to other languages and domains. Further research is needed to explore the adaptation of these models to various languages and understand the unique challenges posed by each language, which is vital for extending the applicability of NER.

IV. Conclusion

The impressive accuracy, precision, recall, and F1-score values demonstrate the model's effectiveness in identifying named entities within job titles and descriptions. These results underscore the NER system's potential to streamline matching job requirements and qualifications, benefiting job seekers and employers.

In addition to quantitative metrics, qualitative assessments were performed by manually inspecting a representative subset of recognized entities—the qualitative analysis aimed to ensure the accuracy and relevance of the named entities identified by the model.

The combination of quantitative and qualitative evaluations provides a comprehensive understanding of the NER system's performance, making it a valuable tool for enhancing job search and talent recruitment processes.

The results presented in this section form the basis for the paper's discussion and conclusion, where the findings' significance is further explored and implications for future work are discussed.

The results of our research shed light on several significant insights and implications for the field of Named Entity Recognition (NER) in the context of Indonesian job vacancies and natural language processing (NLP) in general. This section focuses on analyzing our results' implications, highlighting our research's limitations, and exploring opportunities for future work and innovation. Our Bert-based NER model has demonstrated remarkable accuracy and adaptability in recognizing named entities within job vacancies, contributing to the advancement of NER techniques in the Indonesian job market. The accuracy of 78.5%, precision of 79.7%, recall of 81.1%, and F1 score of 80.8%, respectively, showcase the model's potential practical utility in enhancing job matching and recruitment processes.

However, it's crucial to acknowledge that our model's performance may vary across different sub-domains of job vacancies or specific industries, as it primarily relies on the data it was trained on. Future research may involve domain-specific fine-tuning or ensemble approaches to enhance its effectiveness in diverse job markets. Moreover, the NER task is sensitive to the quality and diversity of the training data, and ongoing efforts to improve the dataset's quality should be considered for even more accurate entity recognition.

One of the notable limitations of our study is the relatively small dataset for individual named entities in standalone testing. This resulted in overlap between tasks and technical skills and between job placement entities. Furthermore, the entity "tasks" may warrant further investigation as it can disambiguate between other entities and holds potential for more in-depth analysis in future research.

While our study marks a significant step forward, it also raises important questions about the generalizability of Bert-based models to other languages and domains. Further research is needed to explore the adaptation of these models to various languages and understand the unique challenges posed by each language, which is vital for extending the applicability of NER.

References

[1] K. R. Chowdhary, "Natural Language Processing," in Fundamentals of Artificial Intelligence. Springer India, 2021. Accessed: Oct. 20, 2024.

[2] N. Nurchim, N. Nurmalitasari, and Z. A. Long, "Indonesian news classification application with named entity recognition approach," JURNAL INFOTEL, vol. 15, no. 2, pp. 130–134, May 2023, doi: 10.20895/infotel.v15i2.909.

[3] S. H. E* and M. A E, "Differential Hiring using a Combination of NER and Word Embedding," International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 1, pp. 1344–1349, May 2020, doi: 10.35940/ijrte.A2400.059120.

[4] F. Stollenwerk, A. Sweden Niklas Fastlund, and A. Nyqvist, "Annotated Job Ads with Named Entity Recognition.", doi: 10.1109/CSCWD49262.2021.9437789.

[5] M. Melih Mutlu and A. Özgür, "A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts.", doi: 10.48550/arXiv.2205.04185.

[6] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," 2020.

DOI : https://doi.org/10.25139/inform.v10i1.7367

[7] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," Aug. 01, 2018, Elsevier Ireland Ltd. doi: 10.1016/j.cosrev.2018.06.001.
[8] J.-J. Decorte, J. Van Hautte, T. Demeester, and C. Develder, "JobBERT: Understanding Job Titles through Skills."
[9] Z. Mincheva, N. Vasilev, V. Nikolov, and A. Antonov, "Extracting Structured Data from Text in Natural Language," International Journal of Intelligent Information Systems, vol. 10, no. 4, p. 74, 2021, doi: 10.11648/j.ijiis.20211004.16.
[10] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," IEEE Trans. Knowl. Data Eng., vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.
[11] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," Computer Science Review, vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.
[12] J.-J. Decorte, J. Van Hautte, T. Demeester, and C. Develder, "JobBERT: Understanding Job Titles through Skills." arXiv, Sep. 20, 2021. doi: 10.48550/arXiv.2109.09605.
[13] H. H. Putro and N. R. Rakhmawati, "Job Standard Parameters from Online Job Vacancy," IJPS, vol. 0, no. 6, p. 46, Mar. 2021, doi: 10.12962/j23546026.y2020i6.8905.

DOI : https://doi.org/10.25139/inform.v10i1.7367