

RESEARCH ARTICLE

Comparative Analysis of Transformer-Based Models for Extracting Skills from Information Technology Job Posting using Named Entity Recognition

Anucha Ruangsiriwattanakul¹, Phamuwat Khanja², Chanida Ruangsiriwattanakul³, and Krit Chaiwannakoo^{1,*}

¹Department of Computer science, Faculty of Science and technology, Uttaradit Rajabhat University, Uttaradit 53000, Thailand

²Department of Computer Engineering, Industrial technology, Uttaradit Rajabhat University, Uttaradit 53000, Thailand

³Department of Information Technology, Faculty of Science and technology, Uttaradit Rajabhat University, Uttaradit 53000, Thailand

*Corresponding author: Krit Chaiwannakoo, krit_chaiwannakoo@uru.ac.th

Article Information

Article History:

Received: 9 June 2025

Revised: 7 October 2025

Accepted: 16 October 2025

Published: 22 June 2026

Keywords:

ALBERT

BERT

DistilBERT

Named Entity Recognition

Natural Language Processing

RoBERTa

Skill Extraction

Transformer-Based Models

XLM-RoBERTa

Abstract

Skill extraction from job postings is critical for labor market analytics and curriculum design, yet it remains challenging due to unstructured text and diverse terminology. This study conducts a comparative evaluation of five Transformer-based Named Entity Recognition (NER) models—BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa—using a curated bilingual dataset of information technology job postings annotated with hard and soft skill entities. The methodology involved standardized preprocessing, BIO tagging, and a 70/15/15 train-validation-test split, with model fine-tuning carried out using optimized hyperparameters. Evaluation metrics included precision, recall, and F1-score, alongside computational efficiency measures such as training time, inference speed, and memory usage. Experimental results indicate that BERT achieved the highest F1-score (0.90–0.91), while DistilBERT delivered a favorable balance between accuracy and efficiency, and ALBERT offered moderate performance with reduced resource demands. In contrast, RoBERTa and XLM-RoBERTa underperformed in this domain-specific context. The contributions of this work include establishing a systematic benchmark for Transformer-based NER in IT skill extraction, providing insights into the trade-offs between accuracy and efficiency, and demonstrating practical implications for human resource analytics and curriculum mapping. These findings advance the understanding of model suitability for domain-specific and bilingual skill extraction tasks and lay the foundation for future cross-lingual and cross-domain applications.

1. Introduction

The rapid growth of the digital economy has amplified the demand for specialized information technology (IT) skills, making online job postings an indispensable resource for labor market intelligence and workforce planning (Adnan and Akbar, 2019; Napierała, 2024; Rahhal et al., 2024). However, extracting skills from job postings remains a non-trivial challenge due to the unstructured nature of text, heterogeneous formats, evolving terminology, and the coexistence of hard and soft skills (Alonso et al., 2025; Kavargyris et al., 2025; Panzaru and Grama, 2025). Traditional manual methods of an-

alyzing such postings are not scalable, highlighting the necessity for automated approaches powered by Natural Language Processing (NLP).

Named Entity Recognition (NER) has emerged as a core NLP task for extracting structured knowledge from unstructured text. With the advent of Transformer-based architectures (Berragan et al., 2023; Košprdić et al., 2024; Gao et al., 2024), such as BERT, RoBERTa, DistilBERT, and ALBERT, NER performance has improved significantly across domains, including healthcare, law, and finance (Shishehgharkhaneh et al., 2024). Recent studies also highlight the growing application of Transformer-based models for labor market analyt-

ics, including skill extraction from job advertisements. Despite these advances, prior research often focuses on monolingual datasets or evaluates models in isolation, leaving a lack of systematic comparative evidence across multiple Transformer variants tailored to IT job postings. This limitation is critical, as model selection directly impacts both extraction accuracy and computational feasibility in real-world HR and educational applications.

Moreover, recent works emphasize that lightweight models such as DistilBERT can deliver efficiency without substantial accuracy loss, while multilingual models such as XLM-RoBERTa show promise for cross-lingual tasks but require domain adaptation for reliable deployment. Nonetheless, comparative evaluations of these models on bilingual IT job postings remain scarce. This gap is especially important in contexts such as South-east Asia, where labor markets are inherently multilingual and domain-specific terminology overlaps between English and local languages.

This study addresses these gaps by providing a systematic comparative analysis of five Transformer-based NER models: BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa for extracting skills from IT job postings. The contributions of this study are fourfold:

1. Construction of a curated bilingual dataset of IT job postings annotated with hard-skill and soft-skill entities.
2. Benchmarking of five Transformer-based NER models under identical conditions, with evaluation based on precision, recall, F1-score, and computational efficiency.
3. Error and efficiency analyses explaining why specific models perform better or worse in IT skill extraction (Fareri et al., 2021).
4. Demonstration of practical applications for curriculum design and workforce analytics, highlighting how NLP-based skill extraction can guide both industry hiring and academic program development.

By clarifying the problem, incorporating recent literature, and presenting the contributions clearly, this study strengthens the novelty and relevance of the research for NLP and labor market analytics.

2. Related Work

2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP) that aims to detect and classify entities in unstructured text into predefined categories such as person, organization, or skill.

Traditional approaches, such as Conditional Random Fields (CRF) and BiLSTM-CRF architectures, demonstrated promising results on structured datasets but struggled with domain-specific and context-dependent texts. With the rise of Transformer-based architectures (Berragan et al., 2023; Gao et al., 2024; Košprdić et al., 2024; Landolsi et al., 2023; Padmanandam et al., 2024), state-of-the-art performance has been achieved across multiple NER benchmarks.

Recent works emphasize the adaptation of Transformer-based NER models to specialized domains, such as biomedical texts, legal documents, and labor market analytics (Landolsi et al., 2023; Shishehgarkhaneh et al., 2024; Rahhal et al., 2024). However, fewer studies have specifically examined skill extraction from job postings in the information technology domain, where entity boundaries are ambiguous and technical terms often overlap with general vocabulary. This highlights a gap that our research seeks to address.

2.2 Transformer-based Models for NER

Transformer-based pre-trained language models, particularly BERT and its variants, have substantially advanced NER tasks. Table 1 summarizes their main characteristics, comparative strengths, and limitations as reported in prior studies (Fareri et al., 2021; Gavrilescu et al., 2025; Gonzalez-Gomez et al., 2025; Tentua et al., 2024; Ternikov, 2022).

Comparative analyses indicate that RoBERTa generally outperforms BERT in accuracy, while DistilBERT offers efficiency with minimal performance loss, making it suitable for low-resource or real-time applications (García Subies et al., 2024; Prenner and Robbes, 2022; Shishehgarkhaneh et al., 2024; Silva Barbon and Akabane, 2022; Zhang et al., 2024). ALBERT, while lightweight, often sacrifices fine-grained recognition, which is critical in extracting specific IT skills. These trade-offs highlight the importance of evaluating models not only by accuracy but also by efficiency and applicability in real-world HR analytics systems.

2.3 Skill Extraction from Job Postings

Skill extraction research has gained momentum as labor market data becomes increasingly digitized. Early approaches relied on dictionary-based methods or statistical co-occurrence models, but these approaches suffered from limited adaptability to evolving skill taxonomies. Recent studies apply Transformer-based NER to automatically extract skills from bilingual or domain-specific corpora (Fareri et al., 2021; Gnehm et al., 2022; Tentua et al., 2024; Akkasi, 2024). For example, Fareri et al. (2021) benchmarked BERT and RoBERTa for skill extraction, reporting significant performance gains over classical models. Recent studies have also shown that Transformer-based and lightweight models can support job description parsing and skill extraction in practical

Table 1. Main characteristics, comparative strengths, and limitations of Transformer-based models reported in prior NER studies.

Model	Key Features	Reported Performance in NER	Limitations
BERT	Bidirectional contextual embeddings; trained on Wikipedia and BookCorpus	Strong baseline across general-domain NER (F1 > 90% on CoNLL-2003)	High computational cost; domain adaptation required
RoBERTa	Optimized training (larger corpus, dynamic masking); robust across domains	Outperforms BERT in large-scale NER benchmarks (Xu et al., 2022)	Requires substantial resources; slower inference
DistilBERT	Compressed version of BERT with approximately 40% fewer parameters	Achieves approximately 97% of BERT’s F1 score with faster training and inference (Zhang et al., 2024)	Slightly lower accuracy in complex, domain-specific NER tasks
ALBERT	Parameter sharing; embedding factorization	Strong efficiency gains; competitive with BERT in some NER tasks (Nguyen et al., 2024)	Struggles with fine-grained entity boundaries
Domain-specific BERT variants (e.g., SciBERT, BioBERT, FinBERT)	Pre-trained on domain-specific corpora	Show improved performance in domain adaptation (Gururangan et al., 2020; Kim et al., 2023)	Require large domain-specific datasets; less generalizable

recruitment contexts (Akkasi, 2024; Gavrilesu et al., 2025).

Despite these advances, most existing studies remain monolingual or domain-specific and lack comparative evaluations across multiple Transformer variants in the IT domain. Addressing this gap, our study provides a systematic comparison of BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa on IT job postings, focusing on both extraction accuracy and computational efficiency.

3. Materials and Methods

This study employs an experimental approach to compare the performance of pre-trained Transformer models for Named Entity Recognition (NER)-based skill extraction from job posting texts (Akkasi, 2024). The research process, illustrated in Fig. 1, comprises five key steps.

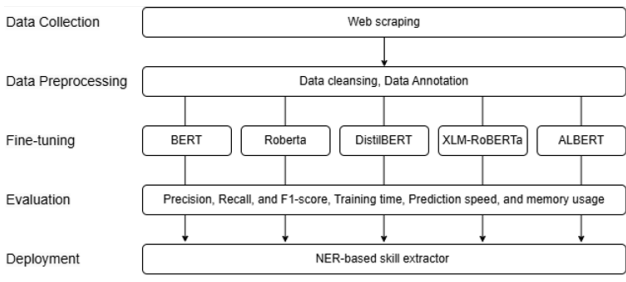


Figure 1. The Research Workflow for Transformer-Based NER Model Comparison.

3.1 Data Collection

IT job postings were collected using web scraping of well-known job search websites in Thailand during the last 6 months of 2024. After filtering the job postings, a total job was found. These were then refined high-quality jobs for further analysis (Fareri et al., 2021; Tentua et al., 2024).

3.2 Data Preprocessing and Annotation

The dataset was cleaned and annotated using the BIO tagging scheme to identify hard-skill and soft-skill entities. Tokenization was applied to each text instance. To guide the annotation process, an expert-defined skill vocabulary list was created. Soft-skill entities are generally more difficult to identify because of ambiguity and contextual variation in job postings.

Table 2 provides representative examples of raw job posting excerpts in English, Thai, and bilingual formats that were subsequently used for the annotation process.

Table 2. Example of annotated dataset entries (BIO tagging format).

ID	Job Title	Raw Job Posting Excerpt
001	Data Scientist (English)	“Looking for candidates proficient in Python, SQL, and Machine Learning. Strong communication skills required.”
002	นักพัฒนาซอฟต์แวร์ (Thai)	“ต้องการผู้สมัครที่มีประสบการณ์ใน จาวา, ไพทอน, และทักษะ การทำงานเป็นทีม.”
003	Business Analyst (Bilingual Thai–Eng)	“เรา require experience in Tableau และความสามารถด้านการสื่อสาร.”

After tokenization and entity annotation, hard-skill and soft-skill entities were labeled according to the BIO tagging scheme.

Table 3 illustrates examples of annotated job posting fragments, highlighting the identification of hard-skill and soft-skill entities in the tokenized text.

Table 3. Example of Tokenized Job Posting with Hard and Soft Skill Annotations.

Sentence Fragment	Annotated Entity	Label
“...experience in SQL and Python for data pipelines”	SQL, Python	Hard Skill
“...strong communication and problem-solving skills”	communication, problem-solving	Soft Skill
“...deploying solutions on AWS and Azure ”	AWS, Azure	Hard Skill

3.3 Fine-tuning and Model Training

Five popular Transformer models were evaluated: BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa. Each model represents a different balance between performance and efficiency:

- BERT provides strong contextual embeddings with bidirectional attention.
- RoBERTa improves upon BERT by using dynamic masking and larger training corpora.
- DistilBERT reduces model size by approximately 40% while retaining most of BERT’s accuracy, enabling faster inference.
- ALBERT employs parameter sharing and factorized embeddings to reduce memory usage.
- XLM-RoBERTa extends RoBERTa to multilingual tasks, making it suitable for bilingual job postings.

The IT job posting dataset was preprocessed through tokenization, lowercasing, and the removal of irrelevant symbols. The data were then divided into 70% for training, 15% for validation, and 15% for testing.

This ratio follows widely accepted practice in NLP experiments, as it provides sufficient data for training while reserving balanced subsets for hyperparameter tuning (validation) and unbiased performance assessment (testing). The choice of 15% for both validation and testing reduces the risk of overfitting while providing reliable estimates of generalization performance, particularly given the heterogeneous nature of job posting texts.

3.4 Evaluation and Comparison

The performance of each model was evaluated using precision, recall, and F1-score metrics derived from the confusion matrix.

- **Precision:** Measures the proportion of predicted entities that are correctly identified. Higher precision indicates fewer false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

- **Recall:** Indicates the model’s ability to identify all relevant skill entities.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

- **F1-score:** Represents the harmonic mean of precision and recall and provides a balanced evaluation when entity classes are unevenly distributed.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Table 4 summarizes the confusion matrix used in this study, which forms the basis for calculating precision, recall, and F1-score in skill entity recognition tasks.

Table 4. Confusion metrics for Skill Entity Recognition.

	Actual Positive	Actual Negative
Predicted Positive	True Positives	False Positives
Predicted Negative	False Negatives	True Negatives

These metrics were selected because NER tasks often require a balance between precision and recall. High precision with low recall indicates that the model can accurately identify skill entities but may fail to detect some relevant skills. Conversely, high recall with low precision indicates that the model can retrieve many skill entities but may incorrectly classify non-skill terms as skills.

3.4.1 Comparative Performance

The performance of the five models was compared in terms of F1-score, error rate, hard-skill and soft-skill discrimination capability, and computational efficiency, including training time, prediction speed, and memory usage. This analysis aims to provide practical guidelines for selecting the most appropriate model for automated IT job market analysis systems.

3.4.2 Error Analysis and Model Limitations

Error Analysis:

An error analysis was conducted to understand why the models misidentified certain skill entities. This analysis examined the types of skills that were most frequently classified incorrectly and investigated how ambiguous language in job postings contributed to these errors. Skill identification from job postings remains challenging because soft-skill entities are often expressed in diverse and context-dependent forms (Fareri et al., 2021; Tentua et al., 2024). Inaccurate skill identification may lead to irrelevant job matching and reduce the effectiveness of recruitment systems (Akkasi, 2024).

Model Limitations:

The limitations of each model were analyzed using both quantitative and qualitative perspectives. Quantitative evaluation was based on confusion-matrix-derived metrics, including precision, recall, and F1-score, while qualitative evaluation focused on common error patterns and challenges observed during experimentation. These evaluation metrics are commonly used in skill identification and information extraction studies (Fareri et al., 2021; Gavrilesu et al., 2025). This analysis provides a comprehensive understanding of the applicability and limitations of each Transformer-based model for skill extraction from job postings.

3.5 Model Deployment

Following pre-training and fine-tuning, the proposed NER-based skill extraction system was deployed to support various applications through REST APIs and a web-based application. This deployment strategy facilitates integration with HR technology platforms and job market analytics systems.

4. Results

This section presents the findings of the study on Transformer model performance for skill extraction.

4.1 Data Collection

The study collected 6,007 IT-related job postings through web scraping from well-known online sources. After cleaning and screening, 3,785 high-quality job postings were retained for model training and evaluation.

4.2 Data Preprocessing and Annotation

A structured data cleaning process was implemented, including the removal of incomplete content and text normalization. Tokenization and BIO tagging enabled accurate classification of hard-skill and soft-skill entities, while careful vocabulary construction ensured con-

sistency in the dataset, which is important for model training.

4.3 Fine-tuning and Model Training

This study presents a comparative evaluation of five pre-trained Transformer models, namely BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa, for skill extraction from IT job postings, focusing on both hard-skill and soft-skill entities. Each model was fine-tuned on a curated dataset that was cleaned, annotated using the BIO tagging scheme, and divided into training, validation, and testing subsets using a 70:15:15 ratio. Fine-tuning employed the AdamW optimizer with controlled hyperparameters, and early stopping was applied to mitigate overfitting. The objective of this process was to adapt the models to the IT job posting domain, particularly to address the ambiguity and context-dependent nature of soft-skill expressions.

Evaluation was conducted using precision, recall, and F1-score, together with computational metrics such as training time and inference speed. The results indicate that BERT and DistilBERT achieved stable convergence and strong performance, whereas RoBERTa and XLM-RoBERTa exhibited slower convergence and lower learning efficiency. Although ALBERT demonstrated rapid initial convergence, it reached a higher final loss than the other models. Model selection was based on validation performance and subsequently confirmed through testing on the held-out dataset. Overall, these findings provide practical insights for selecting appropriate Transformer-based NER models for IT job market analytics while balancing predictive performance and computational efficiency (Fareri et al., 2021; Panzaru and Grama, 2025; Schislyeva and Saychenko, 2022).

Figure 2 presents the training loss comparison of the evaluated Transformer-based models during the fine-tuning process.

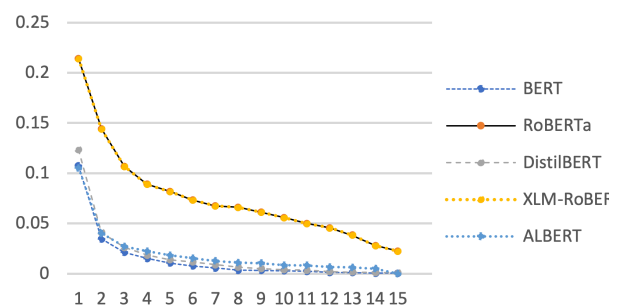


Figure 2. Training loss comparison across Transformer-based models.

4.4 Model Evaluation

The evaluation employed confusion-matrix-based metrics, including precision, recall, and F1-score, to assess

Table 5. Performance comparison of Transformer-based NER on skill entity recognition.

Pretrained	Label	Precision	Recall	F1-score	Support
BERT	HSKILL	0.89	0.91	0.90	3708
	SSKILL	0.91	0.91	0.91	2299
RoBERTa	HSKILL	0.52	0.49	0.50	3681
	SSKILL	0.46	0.42	0.44	2288
DistilBERT	HSKILL	0.88	0.86	0.87	3708
	SSKILL	0.88	0.88	0.88	2299
XLM-RoBERTa	HSKILL	0.52	0.49	0.50	3681
	SSKILL	0.46	0.42	0.44	2288
ALBERT	HSKILL	0.85	0.82	0.84	3708
	SSKILL	0.87	0.88	0.87	2299

each model’s capability in extracting skill-related entities.

Table 5 summarizes the precision, recall, and F1-score achieved by each Transformer-based model for hard-skill and soft-skill entity recognition.

Among the evaluated models, BERT achieved the highest precision, recall, and F1-score for both hard-skill and soft-skill recognition tasks (García Subies et al., 2024; Shishehgharkhaneh et al., 2024; Zhang et al., 2024). DistilBERT delivered comparable performance despite its reduced architecture (Rahhal et al., 2024; Silva Barbon and Akabane, 2022). BERT’s superior performance can be attributed to its bidirectional encoder representations and extensive pretraining on large and diverse corpora, enabling the generation of rich contextual embeddings. Furthermore, its larger hidden representation size and greater number of attention heads enhance the model’s ability to capture long-range dependencies and subtle semantic relationships within skill-related expressions. These characteristics contribute to improved identification of both hard-skill and soft-skill entities (Alonso et al., 2025; Gonzalez-Gomez et al., 2025; Kavargyris et al., 2025).

These findings are consistent with recent NER studies, which reported that BERT frequently outperforms lightweight and multilingual variants in domain-specific entity recognition tasks (García Subies et al., 2024; Zhang et al., 2024). In contrast, both RoBERTa and XLM-RoBERTa exhibited substantially lower performance, which may indicate insufficient domain alignment during the fine-tuning process.

Table 6. Comparative Computational Efficiency of Transformer-Based NER Models.

Model	Training Time (mins/epoch)	Inference Speed (tokens/sec)	Memory Usage (GB)	Computational Efficiency Rating
BERT	45	1203	16	Moderate
DistilBERT	22	2201	8	High
ALBERT	31	1805	10	Moderate
RoBERTa	51	906	18	Low
XLM-RoBERTa	55	802	20	Very Low

Table 6 presents the comparative computational ef-

iciency of the evaluated Transformer-based NER models. The qualitative ratings (“High,” “Moderate,” “Low,” and “Very Low”) were assigned based on relative comparisons of training time, inference speed, and memory usage. Specifically, “High” indicates superior efficiency characterized by low resource consumption and fast inference (e.g., DistilBERT). “Moderate” denotes a balance between computational efficiency and predictive performance with reasonable resource requirements (e.g., BERT and ALBERT). “Low” indicates slower training and greater memory consumption, whereas “Very Low” represents the least efficient models that require substantial computational resources (e.g., RoBERTa and XLM-RoBERTa).

DistilBERT demonstrated the highest computational efficiency, achieving a favorable balance between processing speed and resource utilization. BERT and ALBERT exhibited moderate efficiency, trading some computational speed for improved predictive performance. In contrast, RoBERTa and XLM-RoBERTa required greater computational resources and exhibited slower processing speeds. These findings suggest that DistilBERT provides a practical solution for deployment scenarios where computational efficiency and inference speed are important considerations.

4.4.1 Comparative Performance Analysis

Table 7 summarizes the comparative performance of the evaluated Transformer-based models for hard-skill and soft-skill recognition, highlighting the trade-offs between predictive accuracy and computational efficiency.

Table 7. Comparative Performance of Transformer Models on Skills Recognition Tasks.

Model	Hard Skills (F1-score)	Soft Skills (F1-score)	Overall Performance
BERT	0.90	0.91	Best Performance
DistilBERT	0.87	0.88	Lightweight Alternative
ALBERT	0.84	0.87	Efficient but Less Accurate
RoBERTa	0.50	0.44	Underperformed
XLM-RoBERTa	0.50	0.44	Underperformed

BERT and DistilBERT demonstrated strong performance and appear well suited for deployment in human resource management and recruitment systems. ALBERT achieved competitive results while requiring fewer computational resources, making it a practical option in resource-constrained environments. In contrast, RoBERTa and XLM-RoBERTa exhibited substantially lower performance and may require further optimization and domain adaptation to achieve satisfactory accuracy for skill recognition tasks.

4.4.2 Error Analysis and Model Limitations

The majority of classification errors occurred in the soft-skill category, primarily because of the ambiguous and context-dependent language commonly found in job postings. For example, expressions such as “good communication skills” may not always be explicitly recognized as skill entities, whereas hard-skill entities often correspond to well-defined technical terms, such as “Python” or “Data Analysis,” making them easier to identify accurately. Furthermore, lower-performing models frequently exhibited difficulties in recognizing multi-word skill expressions and domain-specific terminology, leading to incorrect segmentation or semantic misinterpretation. Examples include phrases such as “Full-stack Development” and “Software Engineering Expertise.”

Previous studies have also highlighted the challenges associated with NER-based skill extraction, particularly when models are required to distinguish between contextually similar expressions and skill-related terminology (Fareri et al., 2021; Tentua et al., 2024). These findings suggest that contextual understanding remains a critical factor in improving the accuracy and robustness of skill extraction systems. Although recent approaches have explored large language models to enhance skill extraction performance, challenges related to domain adaptation, entity ambiguity, and computational cost remain important considerations for practical deployment.

4.5 Model Deployment

To improve accessibility and facilitate practical use, the best-performing Transformer-based NER model was deployed on the Hugging Face platform through a public repository maintained by the author. Hugging Face provides an efficient environment for hosting, versioning, and sharing machine learning models, particularly those developed using the Transformers library. The platform also supports deployment through Hugging Face Inference Endpoints, which provide a secure and scalable solution for serving machine learning models in production environments.

After training, the selected NER model was exported in PyTorch format together with its configuration and tokenizer files. The model was subsequently

uploaded to the Hugging Face Model Hub under the profile “Nucha.” Supporting documentation and meta-data were provided to describe the training dataset, task objectives, and model performance.

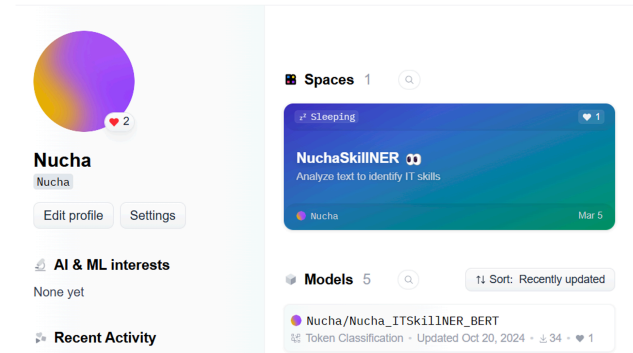


Figure 3. Deployment of the proposed Transformer-based NER model on the Hugging Face platform.

Figure 3 illustrates the deployment of the proposed Transformer-based NER model on the Hugging Face platform, including model hosting and web-based access through Hugging Face Spaces.

Using the Gradio library and Hugging Face Spaces, a user-friendly web interface was developed to improve usability and enable real-time skill entity recognition from raw job descriptions. In addition, the Hugging Face Inference API supports RESTful integration into scalable applications, allowing efficient deployment with batching and inference-speed optimization. Publishing the model as an open-access resource promotes community collaboration, transparency, and research reproducibility. This deployment demonstrates the transition of Transformer-based NER research into a practical AI tool for skill extraction and IT job market analytics.

The Hugging Face platform hosts a wide range of pre-trained models spanning natural language processing, computer vision, and speech recognition tasks. Furthermore, Hugging Face Inference Endpoints facilitate the deployment of machine learning models directly from the Hugging Face Hub, enabling rapid integration into real-world applications.

5. Discussion

Our comparative analysis of five Transformer-based models for skill extraction from IT job postings yields insights directly aligned with the study’s objectives: benchmarking accuracy, assessing efficiency, and identifying practical implications for HR analytics and curriculum design. The results, summarized in Tables 5–7 and Figures fig:fig3–fig:fig4, both confirm established findings and reveal new domain-specific patterns.

BERT’s superior precision confirms its robustness in domain-specific NER. With F1-scores of 0.90 for hard skills and 0.91 for soft skills, BERT outperformed the

other models in identifying IT-related entities. This result is consistent with prior reports in which BERT excelled in specialized domains such as construction supply-chain risk NER (Shishehgharkhaneh *et al.*, 2024). Our findings extend this evidence by demonstrating that, in bilingual IT job postings, BERT’s bidirectional contextual embeddings remain highly effective, surpassing even resource-intensive models such as RoBERTa. This observation suggests that the accuracy advantages reported in other domains can also be transferred to the bilingual recruitment context.

RoBERTa and XLM-RoBERTa underperformed relative to expectations. While earlier studies reported that RoBERTa outperformed BERT in general NLP tasks, our experiments revealed substantially weaker performance, with F1-scores of approximately 0.50. Similarly, XLM-RoBERTa’s multilingual pretraining did not provide a measurable advantage, suggesting that domain alignment is more important than corpus size or multilingual coverage alone. This finding highlights the importance of domain adaptation in skill extraction tasks and demonstrates that model performance cannot be generalized across domains.

DistilBERT achieved performance levels close to those of BERT while maintaining substantially greater computational efficiency. With F1-scores ranging from 0.87 to 0.88 and lower computational requirements, DistilBERT represents a practical solution for real-world HR systems where scalability and cost efficiency are important considerations. This finding directly supports the study’s objective of balancing predictive performance and computational efficiency.

ALBERT demonstrated a trade-off between efficiency and predictive performance. With F1-scores ranging from 0.84 to 0.87, ALBERT provided competitive results while requiring fewer computational resources. However, the model remained less suitable for applications that require highly accurate skill recognition. These findings further clarify the performance limitations of parameter-reduction strategies in skill extraction tasks.

Persistent challenges in soft-skill recognition highlight the complexity of contextual language understanding. Across all models, error analysis revealed greater misclassification rates for soft-skill entities than for hard-skill entities because of the abstract and context-dependent nature of expressions such as “team player” and “good communication.” This observation is consistent with previous studies on skill extraction (Fareri *et al.*, 2021). Furthermore, the bilingual dataset used in this study suggests that linguistic ambiguity may be amplified in multilingual environments, making soft-skill extraction particularly challenging in Southeast Asian recruitment contexts.

By clarifying these confirmations, contradictions, and extensions of previous findings, this study offers three key contributions:

1. Benchmarking both predictive performance and

computational efficiency of Transformer-based models on bilingual IT job postings.

2. Demonstrating the importance of domain alignment over generic multilingual capabilities, challenging assumptions commonly made in general NLP benchmarks.
3. Extending skill extraction research beyond monolingual settings through the use of a Thai–English recruitment corpus, providing new insights into multilingual labor-market analytics.

6. Conclusion

This study presented a comparative analysis of five Transformer-based models, namely BERT, RoBERTa, DistilBERT, ALBERT, and XLM-RoBERTa, for skill extraction from bilingual IT job postings using Named Entity Recognition (NER). The results showed that BERT achieved the highest precision and F1-scores, confirming its robustness in handling domain-specific entities. DistilBERT emerged as a lightweight yet effective alternative, offering competitive performance with substantially lower computational costs. In contrast, RoBERTa and XLM-RoBERTa underperformed, highlighting that strong general-purpose models may not transfer effectively without domain adaptation. ALBERT provided moderate results, balancing efficiency with accuracy, but remained less competitive in precision-critical applications.

Beyond model-level insights, this study highlights persistent challenges in extracting soft-skill entities because of their context-dependent and ambiguous nature, particularly in bilingual job postings. These findings demonstrate both the strengths and limitations of current Transformer architectures and emphasize the importance of adapting models to specific domains and linguistic contexts.

The contributions of this work are threefold: (1) establishing a systematic benchmark of Transformer-based NER models for IT skill extraction, (2) demonstrating the trade-offs between predictive performance and computational efficiency for practical HR and educational applications, and (3) extending skill extraction research to bilingual Thai–English datasets, addressing a gap in the existing literature. These contributions provide useful guidance for researchers, practitioners, and policymakers seeking to integrate AI-driven skill analytics into workforce development and curriculum design.

Future research should expand this comparative framework to additional languages, domains, and real-world recruitment platforms. Incorporating large-scale multilingual corpora and exploring hybrid architectures that combine Transformer-based NER with Large Language Models (LLMs) may further improve performance and scalability. Such directions can contribute to the

development of more inclusive and context-aware labor-market analytics systems.

Acknowledgements

The authors would like to express their sincere gratitude to Uttaradit Rajabhat University for providing financial support for this research. Appreciation is also extended to the organizations and online job-posting platforms that facilitated the data collection and preparation process.

Special thanks are given to the Information Technology (IT) and Human Resources (HR) experts who contributed to skill categorization and the development of the annotation guidelines, thereby improving the accuracy of hard-skill and soft-skill identification.

The authors also acknowledge the Data Engineering team for their support in the development and fine-tuning of the Transformer-based models. In addition, gratitude is extended to the organizations that provided computational resources and infrastructure for model training and experimentation.

CRedit Authorship Contribution Statement

Anucha Ruangsiriwattanakul: Conceptualization, Methodology, Software, Data Curation, Investigation, Visualization, Writing - original draft, Project administration, Funding acquisition. **Phanuwat Khanja:** Software, Formal analysis, Validation, Resources. **Chanida Ruangsiriwattanakul:** Data Curation, Investigation, Validation, Writing - review & editing. **Krit Chaiwannakoop:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of AI Use

This manuscript includes content generated or assisted by artificial intelligence (AI) tools. Specifically, ChatGPT by OpenAI was used to support tasks such as generate sample datasets, language editing, and coding assistance. All AI-generated content was critically reviewed and verified by the authors to ensure accuracy, originality, and compliance with ethical standards.

References

- Adnan, K. and Akbar, R. (2019). An analytical study of information extraction from unstructured and multi-dimensional big data. *Journal of Big Data*, 6(1):91. DOI: 10.1186/s40537-019-0254-8.
- Akkasi, A. (2024). Job description parsing with explainable transformer based ensemble models to extract the technical and non-technical skills. *Natural Language Processing Journal*, 9:100102. DOI: 10.1016/j.nlp.2024.100102.
- Alonso, R., Dessì, D., Meloni, A., and Reforgiato Recupero, D. (2025). A novel approach for job matching and skill recommendation using transformers and the O*NET database. *Big Data Research*, 39:100509. DOI: 10.1016/j.bdr.2025.100509.
- Berragan, C., Singleton, A., Calafiore, A., and Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37(4):747–766. DOI: 10.1080/13658816.2022.2133125.
- Fareri, S., Melluso, N., Chiarello, F., and Fantoni, G. (2021). SkillNER: Mining and mapping soft skills from any text. *Expert Systems with Applications*, 184:115544. DOI: 10.1016/j.eswa.2021.115544.
- Gao, F., Zhang, L., Wang, W., Zhang, B., Liu, W., Zhang, J., and Xie, L. (2024). Named entity recognition for equipment fault diagnosis based on RoBERTa-wwm-ext and deep learning integration. *Electronics*, 13(19):3935. DOI: 10.3390/electronics13193935.
- García Subies, G., Barbero Jiménez, Á., and Martínez Fernández, P. (2024). A comparative analysis of spanish clinical encoder-based models on NER and classification tasks. *Journal of the American Medical Informatics Association*, 31(9):2137–2146. DOI: 10.1093/jamia/ocae054.
- Gavrilescu, M., Leon, F., and Minea, A.-A. (2025). Techniques for transversal skill classification and relevant keyword extraction from job advertisements. *Information*, 16(3):167. DOI: 10.3390/info16030167.
- Gnehm, A.-S., Bühlmann, E., Buchs, H., and Clematide, S. (2022). Fine-grained extraction and classification of skill requirements in german-speaking job ads. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 14–24. Association for Computational Linguistics. DOI: 10.18653/v1/2022.nlpccs-1.2.
- Gonzalez-Gomez, L. J., Hernandez-Munoz, S. M., Borja, A., Arana-Salas, F. A., Azofeifa, J. D., Noguez,

- J., and Caratozzolo, P. (2025). Dynamic taxonomy generation for future skills identification using a named entity recognition and relation extraction pipeline. *Frontiers in Artificial Intelligence*, 8:1579998. DOI: 10.3389/frai.2025.1579998.
- Kavargyris, D. C., Georgiou, K., Papaioannou, E., Petrakis, K., Mittas, N., and Angelis, L. (2025). ES-COX: A tool for skill and occupation extraction using LLMs from unstructured text. *Software Impacts*, 25:100772. DOI: 10.1016/j.simpa.2025.100772.
- Košprdić, M., Prodanović, N., Ljajić, A., Bašaragin, B., and Milošević, N. (2024). From zero to hero: Harnessing transformers for biomedical named entity recognition in zero- and few-shot contexts. *Artificial Intelligence in Medicine*, 156:102970. DOI: 10.1016/j.artmed.2024.102970.
- Landolsi, M. Y., Hlaoua, L., and Ben Romdhane, L. (2023). Information extraction from electronic medical documents: State of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516. DOI: 10.1007/s10115-022-01779-1.
- Napierała, J. (2024). Enhancing taxonomy-based extraction: Leveraging information from online community platforms for digital skills demand identification in job ads. *Statistical Journal of the IAOS*, 40(3):591–602. DOI: 10.3233/SJI-230110.
- Padmanandam, K., Sunitha, K., Jafari, B. M., Jafari, A., Zhao, M., and Pitla, N. (2024). Customized named entity recognition using Bert for the social learning management system platform coursenet-working. *Journal of Computer Science*, 20(1):88–95. DOI: 10.3844/jcssp.2024.88.95.
- Panzaru, C. and Grama, A. (2025). Towards explicit soft skills labelling in ESCO through semantic NLP analysis. *Journal for Labour Market Research*, 59:18. DOI: 10.1186/s12651-025-00409-x.
- Prenner, J. A. A. and Robbes, R. (2022). Making the most of small software engineering datasets with modern machine learning. *IEEE Transactions on Software Engineering*, 48(12):5050–5067. DOI: 10.1109/TSE.2021.3135465.
- Rahhal, I., Kassou, I., and Ghogho, M. (2024). Data science for job market analysis: A survey on applications and techniques. *Expert Systems with Applications*, 251:124101. DOI: 10.1016/j.eswa.2024.124101.
- Schislyaeva, E. R. and Saychenko, O. A. (2022). Labor market soft skills in the context of digitalization of the economy. *Social Sciences*, 11(3):91. DOI: 10.3390/socsci11030091.
- Shishehgarkhaneh, M. B., Moehler, R. C., Fang, Y., Hijazi, A. A., and Aboutorab, H. (2024). Transformer-based named entity recognition in construction supply chain risk management in Australia. *IEEE Access*, 12:41829–41851. DOI: 10.1109/ACCESS.2024.3377232.
- Silva Barbon, R. and Akabane, A. T. (2022). Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: A case study. *Sensors*, 22(21):8184. DOI: 10.3390/s22218184.
- Tentua, M. N., Suprpto, and Afiahayati (2024). NER-Skill.Id: Annotated dataset of indonesian’s skill entity recognition. *Data in Brief*, 53:110192. DOI: 10.1016/j.dib.2024.110192.
- Ternikov, A. (2022). Soft and hard skills identification: Insights from IT job advertisements in the CIS region. *PeerJ Computer Science*, 8:e946. DOI: 10.7717/peerjcs.946.
- Zhang, M., van der Goot, R., Kan, M.-Y., and Plank, B. (2024). NNOSE: Nearest neighbor occupational skill extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 589–608. Association for Computational Linguistics. DOI: 10.18653/v1/2024.eacl-long.35.