BUKTI KORESPONDENSI

Judul :
NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition

1. **Submit Artikel (25 Desember 2023)**
   - **Cover Letter**
   - **Artikel yang di-submit pada sistem**

2. **Review dari DIB (16 Januari 2024)**
   - **Respon dari review**
   - **Revisi pada artikel**

3. **Artikel Accepted (7 Februari 2024)**

4. **Revisi pada sistem DIB telah diterima (18 Februari 2024)**

5. **Pemberitahuan Publish artikel (20 Februari 2024)**

1. **Submit Artikel (25 Desember 2023)**
   - Cover Letter
   - Artikel yang di-submit pada sistem

Gmail

Meilany Nonsi Tentua <meilany@upy.ac.id>

Please verify your contribution to NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition

1 message

Data in Brief <em@editorialmanager.com> Mon, Dec 25, 2023 at 9:46 AM
Reply-To: Data in Brief <dib@elsevier.com>
To: Meilany Nonsi Tentua <meilany@upy.ac.id>

This is an automated message.

Journal: Data in Brief
Title: NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition
Corresponding Author: DR Suprapto Suprapto
Co-Authors: Meilany Nonsi Tentua; Suprapto ; Afiahayati
Manuscript Number: DIB-D-23-02470

Dear Meilany Nonsi Tentua,

The corresponding author DR Suprapto Suprapto has listed you as a contributing author of the following submission via Elsevier's online submission system for Data in Brief.

Submission Title: NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition

Elsevier asks all authors to verify their co-authorship by confirming agreement to publish this article if it is accepted for publication.

Please read the following statement and confirm your agreement by clicking on this link: Yes, I am affiliated.

I irrevocably authorize and grant my full consent to the corresponding author of the manuscript to: (1) enter into an exclusive publishing agreement with Elsevier on my behalf (or, if the article is to be published under a CC BY license, a non-exclusive publishing agreement), in the relevant form set out at www.elsevier.com/copyright ; and (2) unless I am a US government employee, to transfer my copyright or grant an exclusive license of rights (or for CC BY articles a non-exclusive license of rights) to Elsevier as part of that publishing agreement , effective on acceptance of the article for publication. If the article is a work made for hire, I am authorized to confirm this on behalf of my employer. I agree that the copyright status selected by the corresponding author for the article if it is accepted for publication shall apply and that this agreement is subject to the governing law of the country in which the journal owner is located.

If you did not co-author this submission, please contact the corresponding author directly at sprapto@ugm.ac.id.

Thank you,
Data in Brief

More information and support
FAQ: What is copyright co-author verification?
https://service.elsevier.com/app/answers/detail/a_id/28460/supporthub/publishing/

You will find information relevant for you as an author on Elsevier's Author Hub: https://www.elsevier.com/authors
FAQ: How can I reset a forgotten password?
https://service.elsevier.com/app/answers/detail/a_id/28452/supporthub/publishing/kw/editorial+manager/

For further assistance, please visit our customer service site: https://service.elsevier.com/app/home/supporthub/publishing/. Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/7 to our customer support team by phone and 24/7 by live chat and email.

Data in Brief

NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition

--Manuscript Draft--

Manuscript Number:	
Article Type:	Data Article
Keywords:	Natural Language Processing; Named Entity Recognition; Text Mining; Skill Entity Recognition; Indinesian Skill Entity
Corresponding Author:	Suprapto Suprapto Gadjah Mada University INDONESIA
First Author:	Meilany Nonsi Tentua
Order of Authors:	Meilany Nonsi Tentua
	Suprapto Suprapto
	Suprapto
	Afiahayati
Abstract:	NERSkill.Id is a manually annotated named entity recognition (NER) dataset focused on skill entities in the Indonesian language. The dataset comprises 418.868 tokens, each accompanied by corresponding tags following the BIO scheme. Notably, 15,51% of these tokens represent named entities, falling into three distinct categories: hard skill, soft skill, and technology. To construct this dataset, data were gathered from a job portal and subsequently processed using open-source libraries. Given the scarcity of annotated corpora for Indonesian, NERSkill.Id fills a significant void and offers immense value to multiple stakeholders. NLP researchers can harness the dataset's richness to advance skill entity recognition technology in the Indonesian language. Companies and recruiters can benefit by employing NERSkill.Id to enhance talent acquisition and job matching processes through accurate skill identification. Furthermore, educational institutions can leverage the dataset to adapt their courses and training programs to meet the evolving needs of the job market. This dataset can be effectively utilized for training and evaluating named entity recognition systems, empowering advancements in skill entity recognition for the Indonesian language.
Suggested Reviewers:	Kusrini Kusrini kusrini@amikom.ac.id
	Ayu Purwarianti ayu@stei.ac.id
	Tenia Wahyuningrum tenia@ittelkom-pwt.ac.id

Cover letter

Desember 25, 2023

Editorial Data in Brief

Subject: Article Submission - " NERSkill.Id : Annotated Dataset of Indonesian's Skill Entity Recognition "

Dear Editor of Data In Brief

I am submitting a manuscript for consideration of publication in Data in Brief. The manuscript is entitled "**NERSkill.Id : Annotated Dataset of Indonesian's Skill Entity Recognition**". It has not been published elsewhere and that it has not been submitted simultaneously for publication elsewhere.

NERSkill.Id represents a significant milestone in linguistic research, being the initial annotated corpus specifically designed for NER datasets in the Indonesian language. The dedicated focus on skill entities adds a unique dimension to this corpus, addressing a crucial aspect often underrepresented in traditional NLP resources. The importance of this contribution cannot be overstated, as it not only enriches the existing resources for NLP in Indonesian but also opens new avenues for the development of sophisticated language models. I believe that the insights gained from my research on NERSkill.Id would be of great interest to the readership of Data in Brief.

Thank you for considering my submission. I look forward to the opportunity for further discussion and potential publication in Data in Brief.

Yours Sincerely,

Suprapto, Drs., M.I.Kom., Dr.
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences
Universitas Gadjah Mada

Article information

Article title
NERSkill.Id : Annotated Dataset of Indonesian's Skill Entity Recognition

Authors
Meilany Nonsi Tentua[a] Suprapto[b*] Afiahayati[b]

Affiliations
[a]Informatic, Sains and Technology Faculty, Universitas PGRI Yogyakarta, Indonesia
[b]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia.

Corresponding author's email address
sprapto@ugm.ac.id

Keywords
Natural Language Processing, Named Entity Recognition, Text Mining, Skill Entity Recognition, Indinesian Skill Entity

Abstract

NERSkill.Id is a manually annotated named entity recognition (NER) dataset focused on skill entities in the Indonesian language. The dataset comprises 418.868 tokens, each accompanied by corresponding tags following the BIO scheme. Notably, 15,51% of these tokens represent named entities, falling into three distinct categories: hard skill, soft skill, and technology. To construct this dataset, data were gathered from a job portal and subsequently processed using open-source libraries. Given the scarcity of annotated corpora for Indonesian, NERSkill.Id fills a significant void and offers immense value to multiple stakeholders. NLP researchers can harness the dataset's richness to advance skill entity recognition technology in the Indonesian language. Companies and recruiters can benefit by employing NERSkill.Id to enhance talent acquisition and job matching processes through accurate skill identification. Furthermore, educational institutions can leverage the dataset to adapt their courses and training programs to meet the evolving needs of the job market. This dataset can be effectively utilized for training and evaluating named entity recognition systems, empowering advancements in skill entity recognition for the Indonesian language.

Specifications table

Subject	Data science
Specific subject area	Skill Entity Recognition from job description in Indonesian Language
Type of data	Tabular
How the data were acquired	The most data were programmatically scraped using the Beautifulsoup library for Python. The data cleaned and preprocessed using library in Python
Data format	Raw Standardized
Description of data collection	The dataset was compiled using a combination of automated scraping, processing, and manual annotation techniques. Initially, job descriptions from various job vacancies listed on a job portal were extracted through the use of BeautifulSoup Python library. Subsequently, the gathered text files underwent manual annotation, where undergraduate of Informatics annotators labeled each token with the appropriate tag using a spreadsheet application. The final output was exported in a tabular txt format, following the BIO tagging scheme. Each row in the resulting dataset represents a token along with its corresponding tag, enabling the dataset to be effectively utilized for named entity recognition tasks.
Data source location	The Web
Data accessibility	Public Repository Repository Name: Mendeley Data Data identification number: 10.17632/5s8r9ndfvc.1 Direct URL to data: https://data.mendeley.com/datasets/5s8r9ndfvc/1

Value of the data

- NERSkill.Id is the first annotated corpus for NER dataset focused on skill entities in the Indonesian language. It thus makes a valuable contribution to the available resources for Indonesian Language (NLP).
- This dataset is useful for computer NLP research community, companies, recruiters, and educational institutions
- This dataset can be used to evaluation or training in various tasks of skill recognition for transformer language models on the downstream task of NER.
- This dataset follows the BIO format and can thus be combined with other widely used corpora in standard to train large models.

1. Objective

The primary objective of creating this dataset is to procure a precisely annotated Named Entity Recognition (NER) corpus specifically focused on skill entities in the Indonesian language. Although NERSkill.Id is relatively small in size, it has significant potential for fine-tuning language models. Additionally, it can be effectively combined with larger pre-existing corpora to facilitate the training of more comprehensive and adaptable mixed Indonesian models for various NLP tasks.

2. Data description

Following the processes of scraping, preprocessing, and annotation, the ultimate version of the dataset comprises 418.868 tokens. Notably, 15,51% of these tokens correspond to named entities. Before the annotation (tagging) stage, the sentences outlining job requirements undergo a tokenization process. The dataset categorizes named entities into three distinct classes: hard skill, soft skill, and technology [1]. Subsequently, these tokens are marked using the BIO format [2](which stands for Beginning, Inside and Outside). The distribution of these specific named entities within the dataset is shown in Fig.1.

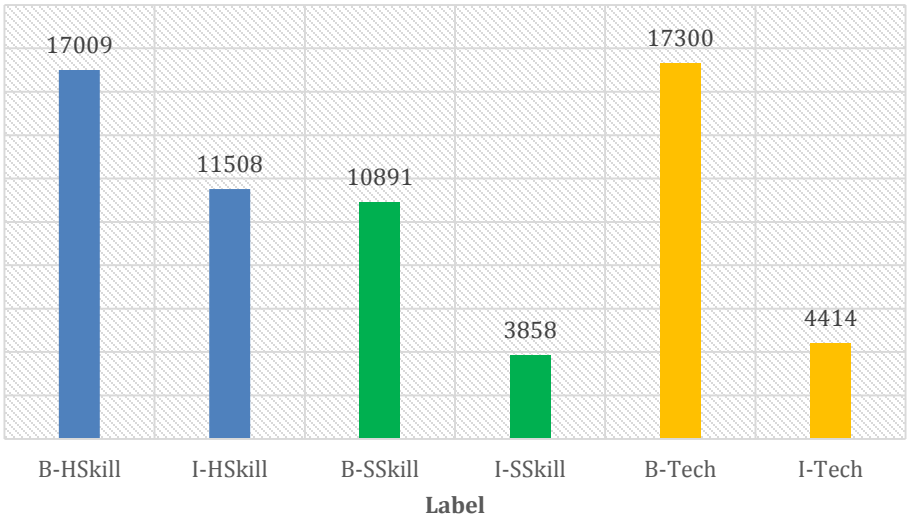


Figure 1. Distribution of Annotation

Hard skill (HSkill) refers to specific abilities required for a job, typically listed under the qualifications section of a job vacancy [3]. Examples of hard skills include web design, computer programming, data analysis, and computer networking. Soft skill (SSkill) encompasses personality traits, personal attributes, and communication abilities needed to interact effectively with others and cultivate sensitivity towards the environment [3]. Examples of soft skills include teamwork, critical thinking, and conflict management. Technology (Tech) represents the type of methods used within Hard Skills [4]. Examples of technologies include C#, Python, MySQL, SQL Server, and Javascript. The annotation table is presented in ConLL2003 format, consisting of 2 columnsword, and tag columns. The NERSkill-ID file is available in .txt format. Table 2 show the description of colomns in NERSkill.Id. Table 3. illustrates the annotation format of the data performed by the NERSkill.ID dataset.

Table 2. Description of columns in NERSkill.Id dataset.

Column	Description
Word	A word, number, or punctuation mark representing one token
Tag	The tag assigned to the token according to the BIO tagging scheme

Table 3. Ilustration of annotation data

Word	Tag
akrab	O
dengan	O
asp.net	B-Tech
core	I-Tech
(c#)	B-Tech
;	O
front-end	B-Hskill
frameworks	I-Hskill

2. Experimental design, materials and methods

Data scraping from job portal. The data used to create the corpus were scraped from the Indeed[1], Jobstreet[2], loker.id[3] dan Job.Id[4]. We used BeautifulSoup as Python library to extract data from indeed and Jobstreet. BeautifulSoup serves as a parser to separate HTML components into a sequence of easily readable elements. We collected manually for job description form loker.id and Job.id. From job portal, 4.394 job description were stored in text files. The full code of data scraping can be found on Mendeley[5].

Data annotation. The text files obtained from the scraping phase were filtered by selecting data with a minimum of 5 words. We divided the files to be annotated into 4 sections. Each file will be annotated manually by 2 different annotators. Eight annotators, all undergraduate informatics students, were employed to annotate skills mentioned in job descriptions using a spreadsheet application. Before distribution the file, the involved annotators convened for a briefing session. The objective was to create a mutual comprehension of the designated tags, which encompassed hard skill, soft skill, and technology. Table 2 shows the annotation rules used for NERSkill.Id. Each sample was collectively deliberated upon, and the author assumed the role of the ultimate decision-maker. Following this, annotations were performed on the annotators' individual computers using a spreadsheet application. In cases of disagreement, the authors intervened to resolve any discrepancies and ensure data quality throughout the

[1] https://id.indeed.com/
[2] https://www.jobstreet.co.id/
[3] https://www.loker.id/
[4] https://job.id/
[5] https://data.mendeley.com/datasets/5s8r9ndfvc/1

annotation process. Once the annotations were finalized, the output file was exported from the spreadsheet in txt format.

Tabel 4. Annotation rules

Entity Entity	Description
B-HSkill	Marks the beginning of a multi-word entity representing a Hard skill
I-HSkill	Refers to the following words within a Hard skill entity after B-HSkill
B-SSkill	Marks the beginning of a multi-word entity representing a Soft skill
I-SSkill	Refers to the following words within a Soft skill entity after B-SSkill
B-Tech	Marks the initiation of a multi-word entity representing a Technology
I-Tech	Refers to the words that follow within a Technology entity after B-Tech
O	Denotes words that do not belong to any recognized entity

Reference results. To test the usefulness of our data in training NER systems, we fine-tuning pretrained model language BERT[5], IndoBERT [6] and EBERT-RP[7] for NER modelling using NERSkill.Id. The model was trained on 5 epochs using a learning rate of 3e-5. The performance of the model on the test set, measured in terms of precision, recall, and F1-score is given in Table 5. We evaluate the model in token level and entity level.

Table 5. Evaluation of reference model on NERSkill.Id

Tag	BERT[5]			IndoBERT[6]			EBERT-RP[7]		
	P	R	F1	P	R	F1	P	R	F1
B-HSkill	84%	89%	87%	83%	88%	85%	88%	92%	90%
B-SSkill	94%	96%	95%	93%	95%	94%	95%	98%	97%
B-Tech	91%	90%	91%	90%	92%	91%	94%	95%	94%
I-HSkill	85%	77%	81%	84%	79%	82%	89%	87%	88%
I-SSkill	90%	90%	90%	94%	91%	93%	93%	86%	90%
I-Tech	74%	69%	72%	77%	66%	71%	88%	76%	82%

*P= Precision; R=Recall; F1=F1-Score

Ethics Statement

The data utilized to construct the dataset do not raise ethical issues, as they were sourced from a Job Portal rather than a social media platform or other sensitive data origins. Permission to employ data from the Job Portal was unnecessary. Our research did not involve any human or animal studies.

Declaration of Competing Interest

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.

Data Availability

NERSkill.Id (Original data) (Mendeley Data).

CRediT author statement

Meilany Nonsi Tentua: Methodology, Software, Investigation, Resources, Data Curation, Writing - Original Draft, Visualization; Suprapto: Investigation, Validation, Writing - Review & Editing, Supervision; Afiahayati: Writing - Review & Editing, Investigation, Validation.

References

[1] CEDEFOP, *Online Job Vacancies and Skills Analysis*. 2019. [Online]. Available: https://www.voced.edu.au/content/ngv:82496

[2] M. Zhang, K. N. Jensen, R. van der Goot, and B. Plank, "Skill Extraction from Job Postings using Weak Supervision," in *CEUR Workshop Proceedings*, 2022, vol. 3218.

[3] Kementrian Ketenagakerjaan and Badan Pusat Statistik, *Klasifikasi Baku Jabatan Indonesia*. Kementrian Ketenagakerjaan dan Badan Pusat Statistik Indonesia, 2014.

[4] ILO, *Indonesia Jobs Outlook 2017: Harnessing Technology for Growth and Job Creation*. 2017.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings ofthe 28th International Conference on Computational Linguistics*, 2021, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[7] M. N. Tentua, Suprapto, and Afiahayati, "An Enhanced Bidirectional Encoder Transformers with Relative Position for Indonesian Skill Recognition," *ICIC Express Lett. - An Int. J. Res. Surv.*, vol. 18, no. In Press., 2024.

Research Data

Click here to download Research Data
https://data.mendeley.com/datasets/5s8r9ndfvc/1

2. Review dari DIB (16 Januari 2024)

- **Respon dari review**
- **Revisi pada artikel**

Gmail

Meilany Nonsi Tentua <meilany@upy.ac.id>

Your Data in Brief Submission: DIB-D-23-02470

1 message

Scientific Editor <em@editorialmanager.com> Tue, Jan 16, 2024 at 8:10 PM
Reply-To: Scientific Editor <dib-me@elsevier.com>
To: Meilany Nonsi Tentua <meilany@upy.ac.id>

You are being carbon copied ("cc:'d") on an e-mail "To" "Suprapto Suprapto" sprapto@ugm.ac.id
CC: "Meilany Nonsi Tentua" meilany@upy.ac.id

Manuscript No.: **DIB-D-23-02470**
Title: NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition
Journal Title: Data in Brief
Corresponding Author: DR Suprapto Suprapto
All Authors: Meilany Nonsi Tentua; Suprapto Suprapto; Suprapto ; Afiahayati
Submit Date: **Dec 24, 2023**

Dear DR Suprapto:

Thank you again for your submission to Data in Brief. Your article will require revision before it can be accepted for publication.

I invite you to revise and resubmit your manuscript after having thoughtfully and carefully addressed the comments below and revising your manuscript accordingly.

I look forward to receiving your revised manuscript by
Jan 31, 2024.

PLEASE NOTE: Please submit your revised manuscript before the given due date as a clean file without comments or tracked changes. Please upload a second version with clear highlights by using the 'Track Changes' function in Microsoft Word, so that changes are easily visible to the editors and reviewers. Please provide a letter to editor to explain point by point the details of the revision and the response to the reviewers' comments. Usually authors are only permitted to revise their article twice for Data in Brief, so carefully address all comments, including formatting requests, when revising your manuscript. If you have any questions, please do not hesitate to contact dib-me@elsevier.com.

Yours sincerely,

Scientific Editor
Data in Brief

Reviewers (if applicable):

Handling editor:
Very clear and quality paper, suitable for the journal and for interesting application in the Name Entity Recognition field.

Just a minor remark noted in page 4: "before distribution the file" --" "before distributing the file"

Scientific editor:
Dear authors,

The handling editor received your manuscript very well, congratulations! There are some minor changes I would like to see addressed before this manuscript can be accepted for publication:
- Please list the address of your institute in the [Data source location] section of the specifications table.
- Data in Brief is a templated journal and does not allow for additional headers to be included. Please remove the Data Availability header at the end of your submission.
- Please provide a Limitations section conform the Data in Brief manuscript template. For more information and instructions, please see the template at the following link http://www.elsevier.com/dib-template.
- Please provide an Acknowledgements section conform Data in Brief's template, including funding sources. For more information and instructions, please see the template at the following link http://www.elsevier.com/dib-template.

I look forward to receiving the revised manuscript. In the meantime, I wish you a good weekend ahead.

With best wishes,

****

*Note: We cannot accommodate PDF manuscript files for production purposes. We also ask that when submitting your revision you follow the journal formatting guidelines. Figures and tables may be embedded within the source file for the submission as long as they are of sufficient

visual quality. For any figure that cannot be embedded within the source file (such as *.PSD, the Photoshop files), the original figure needs to be uploaded separately. Refer to the Guide for Authors for additional information.

To submit your revision, please go to https://www.editorialmanager.com/dib/ and login as an Author.

Your username is: ********

If you need to retrieve your password, please go to:

****

At Elsevier, we want to help all our authors to stay safe when publishing. Please be aware of fraudulent messages requesting money in return for the publication of your paper. If you are publishing open access with Elsevier, bear in mind that we will never request payment before the paper has been accepted. We have prepared some guidelines (https://www.elsevier.com/connect/authors-update/seven-top-tips-on-stopping-apc-scams) that you may find helpful, including a short video on Identifying fake acceptance letters (https://www.youtube.com/watch?v=o5l8thD9XtE). Please remember that you can contact Elsevier s Researcher Support team (https://service.elsevier.com/app/home/supporthub/publishing/) at any time if you have questions about your manuscript, and you can log into Editorial Manager to check the status of your manuscript (https://service.elsevier.com/app/answers/detail/a_id/29155/c/10530/supporthub/publishing/kw/status/).

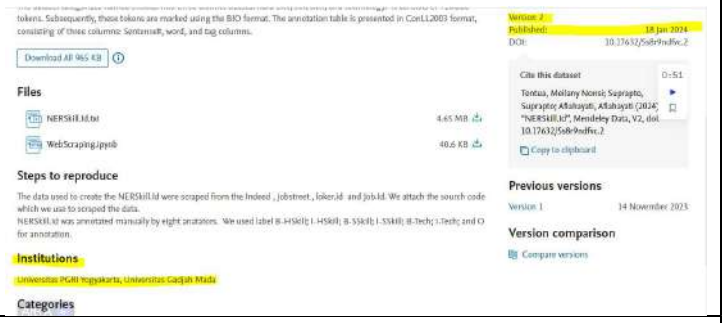
Detailed Response to Editors/ Reviewers

Thank you for the editor's feedback on our manuscript, entitled " **NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition**", and **Manuscript No.: DIB-D-23-02470**. We have revised the manuscript according to the editor's feedback. Here are the details of the changes we have made.

1. Handling editor:

Comment	Respon
Very clear and quality paper, suitable for the journal and for interesting application in the Name Entity Recognition field. Just a minor remark noted in page 4: "before distribution the file" --" "before distributing the file"	Thank you for your kind words and feedback. We have change the word in that sentence. Here is a screenshot of the changes in the manuscript. Before distributing~~on~~ the file,

2. Scientific editor:

Comment	Respon
- Please list the address of your institute in the [Data source location] section of the specifications table.	Thank you for your feedback. We have list the address of our institute in the [Data source location] section of the specifications table. Here is a screenshot of the changes in data source location 
- Data in Brief is a templated journal and does not allow for additional headers to be included. Please remove the Data Availability header at the end of your submission.	Thank you for your feedback. We removed the Data Availability header at the end of your submission.
- Please provide a Limitations section conform the Data in Brief manuscript template. For more information and instructions, please see the template at the following link http://www.elsevier.com/dib-template.	Thank you for your feedback. We provided a Limitations section. We already used the Data in Brief manuscript template as we see at http://www.elsevier.com/dib-template Here is a screenshot of the changes in the manuscript: LIMITATIONS Not applicable

- Please provide an Acknowledgements section conform Data in Brief's template, including funding sources. For more information and instructions, please see the template at the following link http://www.elsevier.com/dib-template.	Thank you for your feedback. We provided an Acknowledgements section conform Data in Brief's template. We have also rearranged the placement as in the template

Data in Brief
Open access
Article template

ARTICLE INFORMATION

Article title

NERSkill.Id : Annotated Dataset of Indonesian's Skill Entity Recognition

Authors

Meilany Nonsi Tentua[a] Suprapto[b*] Afiahayati[b]

Affiliations

[a]Informatic, Sains and Technology Faculty, Universitas PGRI Yogyakarta, Indonesia
[b]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia.

Corresponding author's email address and Twitter handle

sprapto@ugm.ac.id

Keywords

Natural Language Processing, Named Entity Recognition, Text Mining, Skill Entity Recognition, Indonesian Skill Entity

Abstract

NERSkill.Id is a manually annotated named entity recognition (NER) dataset focused on skill entities in the Indonesian language. The dataset comprises 418.868 tokens, each accompanied by corresponding tags following the BIO scheme. Notably, 15,51% of these tokens represent named entities, falling into three distinct categories: hard skill, soft skill, and technology. To construct this dataset, data were gathered from a job portal and subsequently processed using open-source libraries. Given the scarcity of annotated corpora for Indonesian, NERSkill.Id fills a significant void and offers immense value to multiple stakeholders. NLP researchers can harness the dataset's richness to advance skill entity recognition technology in the Indonesian language. Companies and recruiters can benefit by employing NERSkill.Id to enhance talent acquisition and job matching processes through accurate skill identification. Furthermore, educational institutions can leverage the dataset to adapt their courses and training programs to meet the evolving needs of the job market. This dataset can be effectively utilized for training and evaluating named entity recognition systems, empowering advancements in skill entity recognition for the Indonesian language.

SPECIFICATIONS TABLE

Subject	Data science
Specific subject area	Skill Entity Recognition from job description in Indonesian Language
Data format	Raw Standardized
Type of data	Tabular
Data collection	The dataset was compiled using a combination of automated scraping, processing, and manual annotation techniques. Initially, job descriptions from various job vacancies listed on a job portal were extracted through the use of BeautifulSoup Python library. Subsequently, the gathered text files underwent manual annotation, where undergraduate of Informatics annotators labeled each token with the appropriate tag using a spreadsheet application. The final output was exported in a tabular txt format, following the BIO tagging scheme. Each row in the resulting dataset represents a token along with its corresponding tag, enabling the dataset to be effectively utilized for named entity recognition tasks.
Data source location	The Web
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/5s8r9ndfvc.2~~10.17632/5s8r9ndfvc.1~~ Direct URL to data: https://data.mendeley.com/datasets/5s8r9ndfvc/2~~https://data.mendeley.com/datasets/5s8r9ndfvc/1~~

VALUE OF THE DATA

- NERSkill.Id is the first annotated corpus for NER dataset focused on skill entities in the Indonesian language. It thus makes a valuable contribution to the available resources for Indonesian Language (NLP).
- This dataset is useful for computer NLP research community, companies, recruiters, and educational institutions
- This dataset can be used to evaluation or training in various tasks of skill recognition for transformer language models on the downstream task of NER.

- This dataset follows the BIO format and can thus be combined with other widely used corpora in standard to train large models.

BACKGROUND~~OBJECTIVE~~

The primary objective of creating this dataset is to procure a precisely annotated Named Entity Recognition (NER) corpus specifically focused on skill entities in the Indonesian language. Although NERSkill.Id is relatively small in size, it has significant potential for fine-tuning language models. Additionally, it can be effectively combined with larger pre-existing corpora to facilitate the training of more comprehensive and adaptable mixed Indonesian models for various NLP tasks.

DATA DESCRIPTION

Following the processes of scraping, preprocessing, and annotation, the ultimate version of the dataset comprises 418.868 tokens. Notably, 15,51% of these tokens correspond to named entities. Before the annotation (tagging) stage, the sentences outlining job requirements undergo a tokenization process. The dataset categorizes named entities into three distinct classes: hard skill, soft skill, and technology [1]. Subsequently, these tokens are marked using the BIO format [2](which stands for Beginning, Inside and Outside). The distribution of these specific named entities within the dataset is shown in Fig.1.

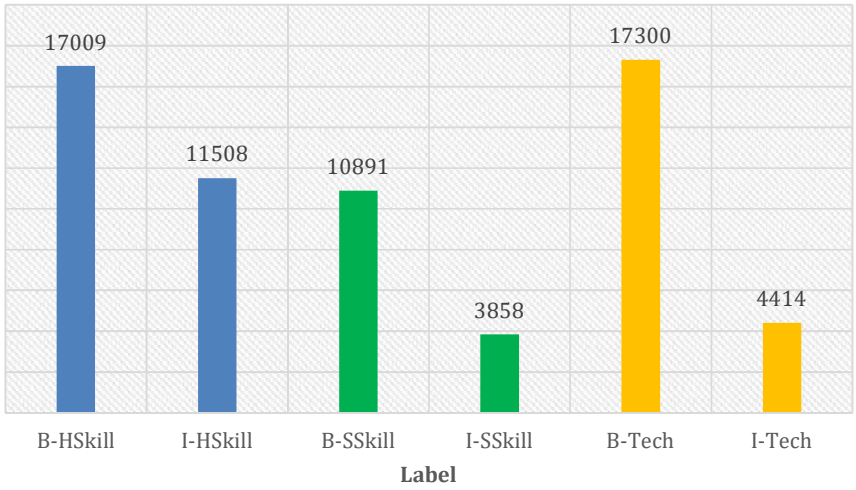


Figure 1. Distribution of Annotation

Hard skill (HSkill) refers to specific abilities required for a job, typically listed under the qualifications section of a job vacancy [3]. Examples of hard skills include web design, computer programming, data analysis, and computer networking. Soft skill (SSkill) encompasses personality traits, personal attributes, and communication abilities needed to interact effectively with others and cultivate sensitivity towards the environment [3]. Examples of soft skills include teamwork, critical thinking, and conflict management. Technology (Tech) represents the type of methods used within Hard Skills

[4]. Examples of technologies include C#, Python, MySQL, SQL Server, and Javascript. The annotation table is presented in ConLL2003 format, consisting of 2 columns word and tag columns. The NERSkill-ID file is available in .txt format. Table 2 show the description of colomns in NERSkill.Id. Table 3. illustrates the annotation format of the data performed by the NERSkill.ID dataset.

Table 2. Description of columns in NERSkill.Id dataset.

Column	Description
Word	A word, number, or punctuation mark representing one token
Tag	The tag assigned to the token according to the BIO tagging scheme

Table 3. Ilustration of annotation data

Word	Tag
akrab	O
dengan	O
asp.net	B-Tech
core	I-Tech
(c#)	B-Tech
;	O
front-end	B-Hskill
frameworks	I-Hskill

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

Data scraping from job portal. The data used to create the corpus were scraped from the Indeed[1], Jobstreet[2], loker.id[3] dan Job.Id[4]. We used BeautifulSoup as Python library to extract data from indeed and Jobstreet. BeautifulSoup serves as a parser to separate HTML components into a sequence of easily readable elements. We collected manually for job description form loker.id and Job.id. From job portal, 4.394 job description were stored in text files. The full code of data scraping can be found on Mendeley Data[5].

Data annotation. The text files obtained from the scraping phase were filtered by selecting data with a minimum of 5 words. We divided the files to be annotated into 4 sections. Each file will be annotated manually by 2 different annotators. Eight annotators, all undergraduate informatics students, were employed to annotate skills mentioned in job descriptions using a spreadsheet application. Before distributingon the file, the involved annotators convened for a briefing session.

[1] https://id.indeed.com/
[2] https://www.jobstreet.co.id/
[3] https://www.loker.id/
[4] https://job.id/
[5] https://data.mendeley.com/datasets/5s8r9ndfvc/2https://data.mendeley.com/datasets/5s8r9ndfvc/1

The objective was to create a mutual comprehension of the designated tags, which encompassed hard skill, soft skill, and technology. Table 2 shows the annotation rules used for NERSkill.Id. Each sample was collectively deliberated upon, and the author assumed the role of the ultimate decision-maker. Following this, annotations were performed on the annotators' individual computers using a spreadsheet application. In cases of disagreement, the authors intervened to resolve any discrepancies and ensure data quality throughout the annotation process. Once the annotations were finalized, the output file was exported from the spreadsheet in txt format.

Tabel 4. Annotation rules

Entity Entity	Description
B-HSkill	Marks the beginning of a multi-word entity representing a Hard skill
I-HSkill	Refers to the following words within a Hard skill entity after B-HSkill
B-SSkill	Marks the beginning of a multi-word entity representing a Soft skill
I-SSkill	Refers to the following words within a Soft skill entity after B-SSkill
B-Tech	Marks the initiation of a multi-word entity representing a Technology
I-Tech	Refers to the words that follow within a Technology entity after B-Tech
O	Denotes words that do not belong to any recognized entity

Reference results. To test the usefulness of our data in training NER systems, we fine-tuning pretrained model language BERT[5], IndoBERT [6] and EBERT-RP[7] for NER modelling using NERSkill.Id. The model was trained on 5 epochs using a learning rate of 3e-5. The performance of the model on the test set, measured in terms of precision, recall, and F1-score is given in Table 5. We evaluate the model in token level and entity level.

Table 5. Evaluation of reference model on NERSkill.Id

Tag	BERT[5]			IndoBERT[6]			EBERT-RP[7]		
	P	R	F1	P	R	F1	P	R	F1
B-HSkill	84%	89%	87%	83%	88%	85%	88%	92%	90%
B-SSkill	94%	96%	95%	93%	95%	94%	95%	98%	97%
B-Tech	91%	90%	91%	90%	92%	91%	94%	95%	94%
I-HSkill	85%	77%	81%	84%	79%	82%	89%	87%	88%
I-SSkill	90%	90%	90%	94%	91%	93%	93%	86%	90%
I-Tech	74%	69%	72%	77%	66%	71%	88%	76%	82%

*P= Precision; R=Recall; F1=F1-Score

LIMITATIONS

Not applicable

ETHICS STATEMENT

The data utilized to construct the dataset do not raise ethical issues, as they were sourced from a Job Portal rather than a social media platform or other sensitive data origins. Permission to employ data from the Job Portal was unnecessary. Our research did not involve any human or animal studies.

CRediT author statement

Meilany Nonsi Tentua: Methodology, Software, Investigation, Resources, Data Curation, Writing - Original Draft, Visualization; Suprapto: Investigation, Validation, Writing - Review & Editing, Supervision; Afiahayati: Writing - Review & Editing, Investigation, Validation.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING 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.

~~Data Availability~~

~~NERSkill.Id (Original data) (Mendeley Data).~~

~~CRediT author statement~~

~~Meilany Nonsi Tentua: Methodology, Software, Investigation, Resources, Data Curation, Writing - Original Draft, Visualization; Suprapto: Investigation, Validation, Writing - Review & Editing, Supervision; Afiahayati: Writing - Review & Editing, Investigation, Validation.~~

REFERENCES

[1] CEDEFOP, *Online Job Vacancies and Skills Analysis*. 2019. [Online]. Available: https://www.voced.edu.au/content/ngv:82496

[2] M. Zhang, K. N. Jensen, R. van der Goot, and B. Plank, “Skill Extraction from Job Postings using Weak Supervision,” in *CEUR Workshop Proceedings*, 2022, vol. 3218.

[3] Kementrian Ketenagakerjaan and Badan Pusat Statistik, *Klasifikasi Baku Jabatan Indonesia*. Kementrian Ketenagakerjaan dan Badan Pusat Statistik Indonesia, 2014.

[4] ILO, *Indonesia Jobs Outlook 2017: Harnessing Technology for Growth and Job Creation*. 2017.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings ofthe 28th International Conference on Computational Linguistics*, 2021, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[7] M. N. Tentua, Suprapto, and Afiahayati, "An Enhanced Bidirectional Encoder Transformers with Relative Position for Indonesian Skill Recognition," *ICIC Express Lett. - An Int. J. Res. Surv.*, vol. 18, no. In Press., 2024.

3. Artikel Accepted (7 Februari 2024)

Gmail

Meilany Nonsi Tentua <meilany@upy.ac.id>

Data in Brief DIB-D-23-02470R1: Your paper has been Accepted for Publication

1 message

Data in Brief <em@editorialmanager.com> Wed, Feb 7, 2024 at 5:25 PM
Reply-To: Data in Brief <dib@elsevier.com>
To: Meilany Nonsi Tentua <meilany@upy.ac.id>

You are being carbon copied ("cc:'d") on an e-mail "To" "Suprapto Suprapto" sprapto@ugm.ac.id
CC: "Meilany Nonsi Tentua" meilany@upy.ac.id

Manuscript No.: **DIB-D-23-02470R1**
Title: NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition
Journal Title: Data in Brief
Corresponding Author: DR Suprapto Suprapto
All Authors: Meilany Nonsi Tentua; Suprapto Suprapto; Suprapto ; Afiahayati
Submit Date: **Dec 24, 2023**

Dear DR Suprapto:

I am pleased to confirm that your paper "NERSkill.Id: Annotated Dataset of Indonesian's Skill Entity Recognition" has been accepted for publication in Data in Brief.

This journal is fully open access; all articles will be immediately and permanently free for everyone to read and download. To provide Open Access, this journal has a publication fee which needs to be met by the authors or their research funders. You will receive the details on the payment in a few days.

Data in Brief employs the CC-BY license for all of its data articles.

Learn about publishing Open Access in this journal: http://www.elsevier.com/journals/data-in-brief/2352-3409/open-access-journal

I encourage you to consider writing and submitting Data in Brief articles about other datasets you may have that could benefit the community. Learn more here: http://www.journals.elsevier.com/data-in-brief/

Submit your latest data article here: http://www.editorialmanager.com/dib

Yours sincerely,

Scientific Editor

Data in Brief

Comments from the handling Editor and Reviewers (if applicable):

The authors have successfully addressed the reviewer's comments. The manuscript is now ready for publication in Data in Brief. Congratulations!

****
For further assistance, please visit our customer support site at http://help.elsevier.com/app/answers/list/p/7923 Here you can search for solutions on a range of topics, find answers to frequently asked questions and learn more about EM via interactive tutorials. You will also find our 24/7 support contact details should you need any further assistance from one of our customer support representatives.

Have you customized a research method or developed a software as part of your research project? Why not publish that work in addition to your 'Data in Brief' article? Go to the MethodsX(http://www.journals.elsevier.com/methodsx) and SoftwareX (http://www.journals.elsevier.com/softwarex/)websites to learn how you can publish your method customizations and software, making your work searchable, peer reviewed, citable and reproducible.

At Elsevier, we want to help all our authors to stay safe when publishing. Please be aware of fraudulent messages requesting money in return for the publication of your paper. If you are publishing open access with Elsevier, bear in mind that we will never request payment before the paper has been accepted. We have prepared some guidelines (https://www.elsevier.com/connect/authors-update/seven-top-tips-on-stopping-apc-

4. Revisi pada sistem DIB telah diterima (18 Februari 2024)

UGM Mail

MEILANY NONSI TENTUA <meilany.nonsi.tentua@mail.ugm.ac.id>

Fwd: Corrections received - [DIB_110192]

1 message

Suprapto Suprapto <sprapto@ugm.ac.id> Sun, Feb 18, 2024 at 3:59 PM
To: MEILANY NONSI TENTUA <meilany.nonsi.tentua@mail.ugm.ac.id>

---------- Forwarded message ---------
Dari: <optteam@elsevierproofcentral.com>
Date: Min, 18 Feb 2024 pukul 09.52
Subject: Corrections received - [DIB_110192]
To: <sprapto@ugm.ac.id>

This is an automatically generated message. Please do not reply because this mailbox is not monitored.

Dear Dr. Suprapto,

Thank you very much for using the Proof Central application for your article "NERSkill.Id: Annotated dataset of Indonesian's skill entity recognition" in the journal "DIB"

All your corrections have been saved in our system. The PDF summary of your corrections, generated from Proof Central, can be downloaded from the following site for your reference:
https://pcv3-elsevier-live.s3.amazonaws.com/4c750c8494c1d856d439030c199cf6/DIB_110192_edit_report.pdf

To track the status of your article throughout the publication process, please use our article tracking service:

http://authors.elsevier.com/TrackPaper.html?trk_article=DIB110192&trk_surname=

For help with article tracking:
http://support.elsevier.com/app/answers/detail/a_id/90

Kindly note that now we have received your corrections, your article is considered finalised and further amendments are no longer possible.

Please help us enhance Proof Central by completing a quick product survey. Your input is valuable and will improve our platform for all users. If you have already participated in the survey during the proof submission, please feel free to ignore this request. To access the survey, kindly click here: Proof Central Survey

For further assistance, please visit our customer support site at http://support.elsevier.com. Here you can search for solutions on a range of topics. You will also find our 24/7 support contact details should you need any further assistance from one of our customer support representatives.

Yours sincerely,
Elsevier Proof Central team

When you publish in an Elsevier journal your article is widely accessible. All Elsevier journal articles and book chapters are automatically added to Elsevier's SciVerse Science Direct which is used by 16 million researchers. This means that Elsevier helps your research get discovered and ensures that you have the greatest impact with your new article.

www.sciencedirect.com

scams) that you may find helpful, including a short video on Identifying fake acceptance letters (https://www.youtube.com/watch?v=o5l8thD9XtE). Please remember that you can contact Elsevier s Researcher Support team (https://service.elsevier.com/app/home/supporthub/publishing/) at any time if you have questions about your manuscript, and you can log into Editorial Manager to check the status of your manuscript (https://service.elsevier.com/app/answers/detail/a_id/29155/c/10530/supporthub/publishing/kw/status/).

---

5. Pemberitahuan Publish artikel (20 Februari 2024)

Gmail

Meilany Nonsi Tentua <meilany@upy.ac.id>

Share your article [DIB_110192] published in Data in Brief

1 message

Elsevier - Article Status <Article_Status@elsevier.com> Tue, Feb 20, 2024 at 2:20 AM
To: meilany@upy.ac.id

ELSEVIER

Share your article!

Dear Dr Tentua,

As co-author of the article *NERSkill.Id : Annotated Dataset of Indonesian's Skill Entity Recognition*, we are pleased to let you know that the final open access version – containing full bibliographic details – is now available online.

The URL below is a quick and easy way to share your work with colleagues, other co-authors and friends. Anyone clicking on the link will be taken directly to the final version of your article on ScienceDirect.

Your article link:
https://doi.org/10.1016/j.dib.2024.110192

You can also use this link to download a copy of the article for your own archive. It also provides a quick and easy way to share your work with colleagues, co-authors and friends. And you are welcome to add it to your homepage or social media profiles, such as Facebook, Google+, and Twitter. Other ways in which you can use your final article have been determined by your choice of user license.

To find out how else you can share your article visit www.elsevier.com/sharing-articles.

Kind regards,
Elsevier Researcher Support

Increase your article's impact
Our Get Noticed guide contains a range of practical tips and advice to help you maximize visibility of your article.

Publishing Lab
Do you have ideas on how we can improve the author experience? Sign up for the Elsevier Publishing Lab and help us develop our publishing innovations!

Have questions or need assistance?
Please do not reply to this automated message.
For further assistance, please visit our Elsevier Support Center where you can search for solutions on a range of topics and find answers to frequently asked questions.
From here you can also contact our Researcher Support team via 24/7 live chat, email or phone support.