Verification of ID Card using Optical Character Recognition (OCR)

Case Study on Eligibility of Subsidy Recipients at PT PLN (Persero)

Rahmat Kurniawan - 235322303
Master's Program in Informatics
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, West Java, Indonesia
23522303@mahasiswa.itb.ac.id

***Abstrak*— This research aims to compare the performance of two Optical Character Recognition (OCR) models, namely Tesseract OCR and PaddleOCR, in extracting data from Identity Cards (KTP) for verifying household electricity subsidy recipients. The household electricity subsidy is a government program aimed at providing affordable electricity access to underprivileged communities. However, the implementation of this program requires accurate and efficient verification of KTP data to ensure that the subsidy is given to the correct recipients. In this study, an analysis was conducted on the accuracy of KTP data extraction, particularly the National Identity Number (NIK) and Name, using Tesseract OCR and PaddleOCR. The results showed that PaddleOCR performed better, with an average NIK accuracy of 93% and Name accuracy of 86%, while Tesseract OCR only achieved an average NIK accuracy of 64% and Name accuracy of 51%. Although PaddleOCR has a slower extraction duration, its high accuracy makes it a more reliable choice for KTP data verification.**

Keywords*— Optical Character Recognition (OCR), Tesseract OCR, PaddleOCR, KTP verification, household electricity subsidy.

I. Introduction

PT PLN (Perusahaan Listrik Negara) is a State-Owned Enterprise (BUMN) responsible for electricity provision in Indonesia and has been assigned by the government through the Ministry of Energy and Mineral Resources (ESDM) to distribute subsidized household electricity connections to underprivileged communities. The criteria for underprivileged communities are stipulated in the Minister of Energy and Mineral Resources Regulation No. 3 of 2024 and can be served with subsidized household tariff groups with power R1/450 VA or R1/900 VA [1].

PLN's responsibilities include ensuring that electricity user data is recorded in the Integrated Data of the Ministry of Social Affairs (DTKS) and verifying that Identity Cards (KTP) are accurate. The challenge faced is the misuse of user data or KTP by unauthorized parties, which can lead to misdirected subsidy distribution and state losses.

Currently, PLN's system is integrated with the Ministry of Social Affairs through the Ministry of Energy and Mineral Resources to determine the eligibility of subsidized household connections. However, KTP verification has not been carried out during the application process and is instead conducted during the connection process.

Machine learning technology, specifically Optical Character Recognition (OCR), can be used to verify KTPs in less than 5 seconds using the Pytesseract library and image processing [2]. In another case, license plate detection using the PaddleOCR library and YOLOv8 architecture achieved an accuracy of 0.871 in detecting license plate numbers [3]. Additionally, the similarity of a name extracted by OCR with the actual name can be assessed using the Fuzzy Wuzzy library [4].

By utilizing OCR, image processing, and entity matching, it is expected to streamline the KTP verification process, allowing it to be conducted during the application for electricity connections. Moreover, OCR is expected to improve subsidy distribution accuracy and reduce data misuse potential.

This research aims to compare the performance of two OCR libraries, Tesseract OCR and PaddleOCR, using image processing and similarity assessment of extracted results using the Fuzzy Wuzzy library.

II. Related Work

A. Identity Cards (KTP)

The Identity Card (KTP) is the official identification for residents, serving as proof of identity issued by the Ministry of Home Affairs - Directorate General of Population and Civil Registration. The regulations concerning KTP and population administration are outlined in Law No. 24 of 2013 on the Amendment of Law No. 23 of 2006 on Population Administration [5].

In the context of subsidized household connections, the National Identity Number (NIK) is used to search for eligibility information of subsidy recipients as determined by the Ministry of Social Affairs [1]. According to the provisions of the Minister of Energy and Mineral Resources Regulation No. 3 of

2024 on the Provision of Electricity Tariff Subsidies for Household Consumers of PT Perusahaan Listrik Negara (Persero), a subsidized household connection can only be granted to one household [1]. Therefore, PLN must ensure that the recipients of subsidized household connections match the KTP data provided and that there is no misuse of data.

B. Subsidized Household Electricity Connections

Subsidized Household Electricity Connections provided by PLN are regulated under the Minister of Energy and Mineral Resources Regulation No. 3 of 2024 on the Provision of Electricity Tariff Subsidies for Household Consumers of PT Perusahaan Listrik Negara (Persero) [1]. Recipients of subsidized household electricity connections can be served under the R1/450 VA or R1/900 VA tariff groups by meeting the following criteria:

- R1/450 VA tariff group: The NIK is registered in the Integrated Data of the Ministry of Social Affairs (DTKS) or the recipient's village/sub-district location is in the 3T (Frontier, Outermost, and Underdeveloped) areas.
- R1/900 VA tariff group: The NIK is registered in the Integrated Data of the Ministry of Social Affairs (DTKS).

To ensure the accuracy of subsidized household electricity connections, PLN is responsible for verifying the applicant's data against the KTP.

C. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is the process of converting printed or handwritten text images into editable digital computer formats [2]. In several cases, OCR significantly contributes to the development of digital libraries [6]. OCR is a complex software implemented in various fields to automatically read data and enter it into databases. Examples of its use include scanning passports, bank transfer notes, securities, recognizing vehicle license plates from videos or images, and preserving the content of major reference books and historical manuscripts. In this research, the Python libraries used are:

- Tesseract OCR, an open-source software initially developed by Hewlett-Packard between 1985 and 1995, is now developed by Google and released under the Apache license. Tesseract OCR is an automatic pattern recognition tool that supports Unicode (UTF-8) and over 100 languages. In the automatic recognition process, Tesseract OCR compares the input image with pre-defined images to find a match. This tool can be used to recognize various types of patterns, including signatures, fingerprints, images, and even a person's face. Here are some steps in the Tesseract OCR method [3].
- PaddleOCR, an open-source software created by the Chinese company Baidu, uses Differentiable Binarization (DB) for text detection and the CRNN (Convolutional Recurrent Neural Network) model for text recognition. PaddleOCR is a highly efficient and practical Optical Character Recognition (OCR) system developed by researchers from Paddle. This technology is designed to automatically recognize text in images, and PaddleOCR has successfully created very lightweight OCR models without sacrificing performance.

D. Performance Evaluation

The evaluation of text extraction results using OCR will utilize the Fuzzy Wuzzy method. Fuzzy Wuzzy is a technique that uses fuzzy string matching to match entities from two different data sources [7]. This method can handle differences in token order within a string, thus addressing discrepancies in text extraction results from OCR and allowing for corrections.

Additionally, the use of the Fuzzy Wuzzy metric is based on the practices of the Directorate General of Population and Civil Registration - Ministry of Home Affairs in matching and verifying KTP data. This method is currently used in the KTP data verification process. Therefore, this aligns with the matching process that PLN will conduct with the Ministry of Social Affairs.

III. METHODOLOGY

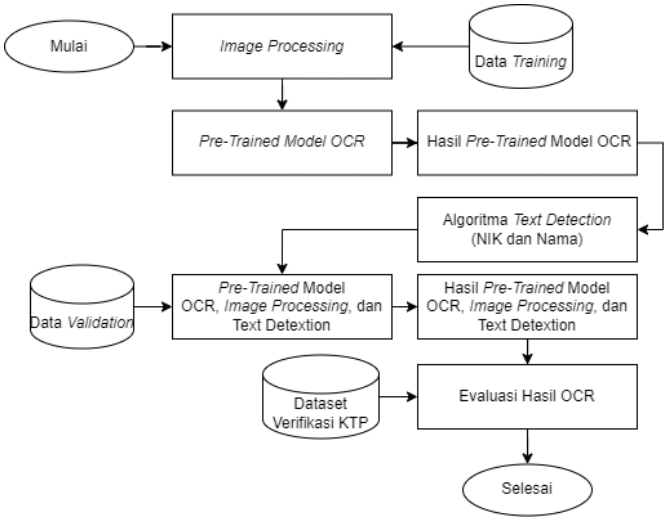

Fig. 1. Rancangan metodologi ekstraksi teks dan deteksi teks

Figure 1 represents the research design. In this study, there are two main processes: text extraction using OCR and the development of an algorithm to detect NIK and Name text. In the text extraction process, pre-trained models from Tesseract OCR and PaddleOCR are used with a testing scenario comparing the performance of both pre-trained models with and without image processing. Subsequently, to obtain the extraction results, an algorithm is developed to detect NIK and Name text.

A. Dataset

In this research, two datasets are used: the KTP dataset and the KTP verification dataset. The KTP dataset used originates from the data of customers receiving subsidized household electricity connections with 450 VA and 900 VA power, submitted in 2023, within the service area of the Customer Service Unit (ULP) Bondowoso - UP3 Situbondo - UID East Java. The data used in this study consists of 91 KTPs along with

a KTP verification dataset containing NIK and Name information as per the KTP. An example of the KTP verification data can be seen in Figure 2.

NAMA FILE	NIK	NAMA	IDPEL	TARIF	DAYA
upl1-383438-17(	351101130(	MUHAMMAD	5165327	R1T	450
upl1-327608580	327608580(	SRI AYU NAFI	5165326	R1T	450
upl1-331602650	331602650:	SITI JARWATI	5165302	R1	450
upl1-331901190	331901190:	ALI ROSYIDI	5165302	R1	450
upl1-350903641	350903641(	ERNAWATI	5165306	R1T	450
upl1-350905460	350905460:	RIA INDRAW/	5165306	R1T	450
upl1-350911570	350911570:	TITIN KUMAL	5165327	R1T	450
upl1-350912500	350912500:	KHOIRIATUL I	5165324	R1T	450

Fig. 2. Dataset Hasil Verifikasi KTP

B. *Preprocessing*

The preprocessing stage is the most crucial step in preparing the dataset for the OCR process. In this research, data preprocessing is performed using several images processing techniques, including image rotation, image cropping, image resizing, and converting the image format to binary.

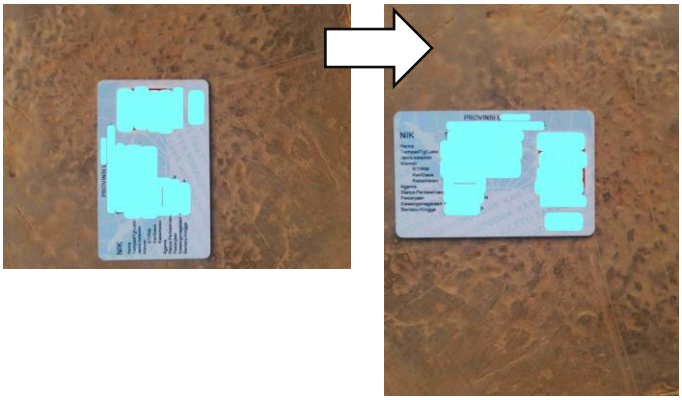

Fig. 3. Image Processing – Image Rotate

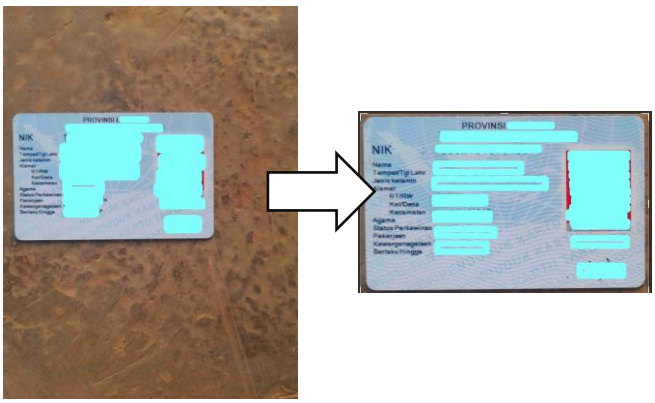

Fig. 4. Image Processing – Image Cropping

Figure 3, illustrates the image rotation technique used to standardize the orientation of KTP images within the dataset. This is important because orientation variations can cause difficulties in model training. Next, the image cropping technique, as seen in Figure 4 image, ensures that the KTP image is clearly visible and does not get compressed during image resizing.

C. *Training Model*

At this stage, the OCR models used are pre-trained models available in Python libraries, specifically Tesseract version 5.3.3.20231005 and PaddleOCR version 2.7.3. The next step involves preparing the necessary dataset and randomly sampling images. The sampled images then undergo image processing. This processing is carried out step-by-step and tested on each pre-trained model. This process must be performed carefully to observe the impact of the image processing. The focus of this training is to enable the pre-trained models to extract the NIK and Name data from the KTP.

The extraction results from the pre-trained models are then used to create an algorithm to retrieve the NIK and Name data. This presents a challenge, as the extraction results from the pre-trained models are not always perfect. Therefore, an approach is needed to effectively extract the NIK and Name data from the extraction results.

IV. Discussion and Results

A. *Discussion*

In this section, we discuss the test results from scenarios conducted using two OCR models: Tesseract OCR and PaddleOCR. Three testing scenarios were carried out to evaluate the performance of both models in extracting text from KTPs. Each scenario had a different approach to using image processing, aimed at improving the accuracy and efficiency of text extraction. The testing scenarios were conducted as follows:

- Without Image Processing: This scenario aimed to evaluate the ability of the pre-trained models to extract text without using any image processing. The experimental results in the Table I show that Tesseract OCR failed to extract NIK and Name features with adequate accuracy. In contrast, PaddleOCR showed better results with 100% accuracy for Name data. This indicates that without image processing, the text extraction capabilities of both models are very limited.

TABLE I. Pre-Trained Model Without Image Processing

No.	Model	Resolution	Time (seconds)	Accuracy (%)
1	*Tesseract OCR*	3000 x 4000	2.99	NIK: 0 Nama: 86
2	*PaddleOCR*	3000 x 4000	5.98	NIK: 0 Nama: 86

- Using Image Processing, In this scenario, we applied various image processing techniques such as image rotation, cropping, resizing, and format conversion. The experimental results in the Table II show a significant improvement in the performance of both models. Tesseract OCR showed an increase in accuracy with various image processing techniques, especially when using image rotation and format conversion, achieving 100% accuracy for NIK and 75% for Name. PaddleOCR demonstrated excellent performance with 100% accuracy for all parameters after applying image processing. This indicates that image processing plays a crucial role in improving text extraction accuracy.

TABLE II. PRE-TRAINED MODEL WITH IMAGE PROCESSING

No.	Model	Image Processing	Resolution	Time (seconds)	Accuracy (%)
1	*Tesseract OCR*	Rotate	3000 x 4000	0,89	NIK: 94 Nama: 100
2	*Tesseract OCR*	Add Cropping	2026 x 1276	2,44	NIK: 0 Nama: 100
3	*Tesseract OCR*	Add Resize	1013 x 638	0,52	NIK: 0 Nama: 0
4	*Tesseract OCR*	Convert to binary	1013 x 638	0,40	NIK: 100 Nama: 100
5	*PaddleOCR*	Rotate	3000 x 4000	3,00	NIK: 100 Nama: 100
6	*PaddleOCR*	Add Cropping	2026 x 1276	3,39	NIK: 100 Nama: 100
7	*PaddleOCR*	Add Resize	1013 x 638	2,85	NIK: 100 Nama: 100
8	*PaddleOCR*	Convert to binary	1013 x 638	3,64	NIK: 100 Nama: 100

- Using Image Processing with NIK and Name Detection, In this scenario, we implemented a feature detection algorithm for NIK and Name on KTPs after applying image processing techniques. The tests were conducted on the entire available KTP dataset. The experimental results in the Table III show that Tesseract OCR has an average accuracy of 64% for NIK and 51% for Name, whereas PaddleOCR achieved 93% accuracy for NIK and 86% for Name. However, Tesseract OCR was more efficient in terms of OCR duration compared to PaddleOCR.

TABLE III. PRE-TRAINED MODEL WITH IMAGE PROCESSING AND FEATURE DETECTION

No.	Model	Image Processing	Number of KTP	Average Time (seconds)	Average Accuracy (%)
1	*Tesseract OCR*	All	91	1,37	NIK: 64 Nama: 51
2	*PaddleOCR*	Rotate, Cropping, Resize	91	4,63	NIK: 93 Nama: 86

B. Results

Based on the experiments conducted, it can be concluded that the use of image processing significantly improves the performance of text extraction from KTPs. Key points derived from these results include:

- Use of Image Processing: Techniques such as image rotation, cropping, resizing, and format conversion have been shown to enhance the text extraction accuracy of both OCR models. This underscores the importance of preprocessing in OCR applications.
- Model Comparison: PaddleOCR consistently demonstrated better performance compared to Tesseract OCR in terms of accuracy. This could be attributed to the more advanced architecture and algorithms employed in PaddleOCR.
- Text Detection Accuracy: Although both models showed improvement after image processing, PaddleOCR excelled in text detection accuracy for critical features like NIK and Name, which are essential for validating subsidy recipients.
- Processing Duration: While PaddleOCR required more time in some cases, its efficiency in terms of accuracy makes it a better choice for applications that require high levels of accuracy.

V. CONCLUSION

These results demonstrate that with the appropriate implementation of image processing techniques, the capabilities of pre-trained OCR models can be significantly enhanced. For applications requiring high accuracy in text extraction from official documents like KTPs, PaddleOCR with image preprocessing is a more effective solution compared to Tesseract OCR. For the implementation of KTP data verification for subsidy recipients, a threshold percentage value for acceptable NIK and Name matches should be established, as extraction results do not always yield perfect scores. Future research can focus on further optimizing image processing techniques and exploring other OCR models that may offer better performance.

ACKNOWLEDGMENTS

Firstly, the author is an employee of PT PLN (Persero) Headquarters, in the Customer Experience and Excellent Services Division (DIV CES). Currently, the author is pursuing a master's degree in the Faculty of Electrical Engineering and Informatics, Bandung Institute of Technology. PT PLN (Persero) has always supported this research.

DECLARATION

I hereby declare that the paper I have written is my own work, not an adaptation or translation of someone else's paper, and is not plagiarized.

Bandung, 05 June 2024

TTD

Rahmat Kurniawan - 23522303

REFERENSI

[1] ESDM, “Permen ESDM Nomor 3 Tahun 2024.” Diakses: 26 April 2024. [Daring]. Tersedia pada: https://jdih.maritim.go.id/permen-esdm-no-3-tahun-2024

[2] F. M. Rusli, K. A. Adhiguna, dan H. Irawan, “Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing,” dalam *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 2021, hlm. 621–626. doi: 10.1109/ICoICT52021.2021.9527510.

[3] L. Satya, M. R. D. Septian, M. W. Sarjono, M. Cahyanti, dan E. R. Swedia, “SISTEM PENDETEKSI PLAT NOMOR POLISI KENDARAAN DENGAN ARSITEKTUR YOLOV8,” *Sebatik*, vol. 27, no. 2, Art. no. 2, Des 2023, doi: 10.46984/sebatik.v27i2.2374.

[4] “FuzzyWuzzy Python library,” GeeksforGeeks. Diakses: 31 Mei 2024. [Daring]. Tersedia pada: https://www.geeksforgeeks.org/fuzzywuzzy-python-library/

[5] Undang-Undang RI, “Perubahan atas Undang-Undang Nomor 23 Tahun 2006 tentang Administrasi Kependudukan,” Database Peraturan | JDIH BPK. Diakses: 4 Juni 2024. [Daring]. Tersedia pada: http://peraturan.bpk.go.id/Details/38985/uu-no-24-tahun-2013

[6] A. L. Firdaus, M. S. Kurnia, T. Shafera, dan W. I. Firdaus, “Implementasi Optical Character Recognition (OCR) Pada Masa Pandemi Covid-19,” *JUPITER: Jurnal Penelitian Ilmu dan Teknologi Komputer*, vol. 13, no. 2, Art. no. 2, Okt 2021, doi: 10.5281/3912.jupiter.2021.10.

[7] I. S. Ardan, M. J. Sulastri, dan N. A. Rakhmawati, “ANALISIS PERFORMANSI ENTITY MATCHING DENGAN FUZZY WUZZY PADA ARTIKEL FAIRNESS AI,” *Jurnal Teknoinfo*, vol. 17, no. 2, Art. no. 2, Jul 2023, doi: 10.33365/jti.v17i2.2711.