Perbandingan Unjuk Kerja *Library Optical Character Recognition* (OCR) dalam Pengenalan Teks pada Dokumen Digital

Muhammad Noko Darpito [1], Kartika Firdausy [2], Abdul Fadlil [3]

[1,3] Magister Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
[2] Program Studi Teknik Elektro, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
[1]2407048001@webmail.uad.ac.id, [2]kartika.firdausy@te.uad.ac.id, [3]fadlil@mti.uad.ac.id

Abstrak

Optical Character Recognition (OCR) merupakan teknologi yang digunakan untuk mengubah teks dalam dokumen digital menjadi teks yang dapat dikenali oleh mesin. Pemilihan metode OCR yang tepat sangat bergantung pada efisiensi pemrosesan dan akurasi pengenalan teks, terutama dalam penerapan yang membutuhkan kecepatan tinggi dan tingkat kesalahan minimal. Dalam penelitian ini, dilakukan perbandingan performa antara *Tesseract* dan *EasyOCR* melalui metode penelitian yang mencakup tahapan pengumpulan data, ekstraksi teks, implementasi OCR menggunakan kedua library tersebut, dan evaluasi hasil ekstraksi teks kedua library OCR tersebut menggunakan *Word Error Rate* (WER), *Character Error Rate* (CER) dan akurasi ekstraksi OCR keseluruhan. Dataset yang digunakan yang terdiri dari 50 dokumen formulir dengan variasi tata letak dan ukuran font, serta 10 dokumen artikel dengan variasi format huruf (standar dan kapital). Hasil penelitian menunjukkan bahwa *Tesseract* secara konsisten lebih cepat dalam memproses dokumen, dengan waktu rata-rata 0,34 detik per dokumen formulir dibandingkan *EasyOCR* yang memerlukan 1,81 detik. Namun, *EasyOCR* memperlihatkan performa yang lebih baik dalam akurasi pengenalan teks, dengan nilai WER rata-rata yang lebih rendah sebesar 25,78% dibandingkan *Tesseract* sebesar 49,69% pada dokumen formulir. Dengan demikian, *Tesseract* lebih sesuai untuk pemrosesan cepat dalam jumlah besar, sedangkan *EasyOCR* lebih direkomendasikan untuk dokumen dengan kompleksitas tinggi yang membutuhkan akurasi lebih baik.

Kata kunci: OCR, *Tesseract*, *EasyOCR*, digitalisasi

1. Pendahuluan

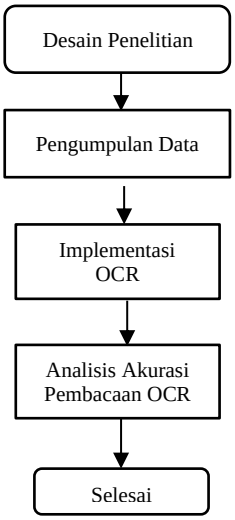
Pada era digital, teknologi *Optical Character Recognition* menjadi sangat penting dalam mengonversi teks dari dokumen cetak ke dalam format digital yang dapat dibaca oleh komputer. *Optical Character Recognition* (OCR) adalah teknologi yang yang memungkinkan komputer untuk mengidentifikasi dan mengubah teks dalam format gambar menjadi data yang dapat diedit dan diproses secara digital (Banu et al., 2023). Dasar dari teori OCR berakar dari beberapa teknik pemrosesan citra dan pembelajaran mesin yang memungkinkan sistem untuk mengenali karakter yang dicetak atau tulisan tangan (Setyadi & Susetyo, 2023). Proses OCR terdiri dari beberapa tahapan, termasuk pra-pemrosesan, segmentasi, ekstraksi fitur, dan klasifikasi, yang masing-masing memiliki peran krusial dalam meningkatkan akurasi dan kecepatan pengenalan karakter (Nguyen et al., 2022).

Tesseract dan *EasyOCR* merupakan dua library OCR yang banyak digunakan untuk pengolahan teks dalam dokumen digital. Tesseract, yang dikembangkan oleh HP *Labs* dan kini disponsori oleh *Google*, telah lama menjadi solusi *open-source* yang populer untuk pengenalan teks. *Tesseract* menggunakan pendekatan berbasis *Long Short-Term Memory* (LSTM) untuk meningkatkan akurasi pengenalan karakter, tetapi masih memiliki tantangan dalam menangani teks yang terdistorsi atau tidak teratur (Sporici et al., 2020). Di sisi lain, *EasyOCR* adalah *library* OCR berbasis *deep learning* yang lebih modern dan menawarkan kemudahan implementasi serta dukungan untuk berbagai bahasa termasuk Bahasa Indonesia (Banu et al., 2023).

Penelitian ini bertujuan untuk membandingkan kinerja *library Tesseract* dan *EasyOCR* dalam mengenali teks pada dokumen digital formulir dan dokumen artikel berdasarkan tiga aspek utama antara lain kecepatan pemrosesan, akurasi pengenalan, dan fleksibilitas dalam menangani berbagai format dokumen (Asroni et al., 2023). Dataset yang digunakan dalam penelitian adalah dokumen formulir dan artikel yang diperoleh dari *platform Scribd*. Dengan menggunakan dataset tersebut yang mencakup variasi dalam tata letak, dan variasi huruf, penelitian ini akan memberikan wawasan lebih lanjut mengenai kelebihan dan kekurangan masing-masing *library* OCR. Hasil penelitian ini diharapkan dapat membantu dalam memilih solusi OCR yang paling sesuai untuk kebutuhan pengolahan dokumen digital dalam berbagai bidang, seperti digitalisasi arsip, otomatisasi administrasi, dan pengelolaan basis data dokumen elektronik (Raswaty & Nuryuliani, 2021) (Hamdi et al., 2021) (Hegghammer, 2022) (Sharma et al., 2023) (Al amin & Aprilino, 2022).

2. Metode

Di dalam penelitian ini dilakukan serangkaian langkah sistematis untuk mengevaluasi kedua *library* OCR *Tesseract* dan *EasyOCR* dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

1. Desain Penelitian

Penelitian ini dirancang untuk mengukur perbedaan kinerja antara *Tesseract* dan *EasyOCR* dalam hal akurasi pengenalan teks dan kecepatan pemrosesan. Tujuan utama penelitian ini adalah mengevaluasi sejauh mana kedua *library* dapat mengenali teks dari dokumen digital yang memiliki format bervariasi, seperti formulir yang mengandung tabel, kolom isian, dan elemen grafis lainnya.

2. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini terdiri dari 50 dokumen digital dalam format *image* (jpg) berbentuk formulir dengan berbagai tingkat kompleksitas tata letak, jumlah karakter, dan ukuran font yang berbeda sebagaimana pada Gambar 2.

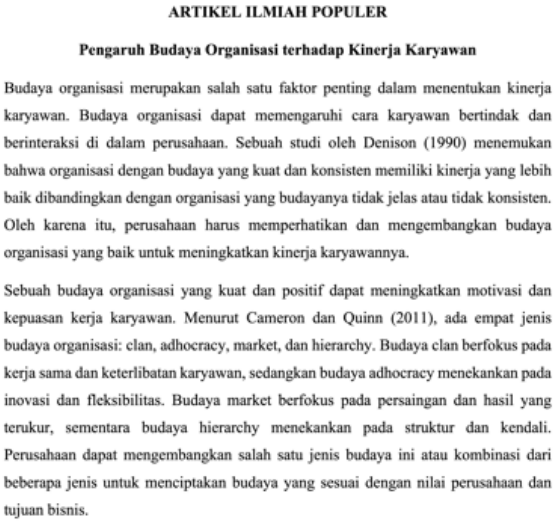
FORMULIR PENDAFTARAN LOMBA
CERDAS CERMAT ASEAN

1. Identitas

A. Nama Lengkap	: Vina Kartika
B. Umur	: 12
C. Kelas	: VI (LIMA)
D. Alamat	: Jl. Melati No.24
Kota	: Surabaya
Kode Pos	: 60482
E. No.Telpon Rumah	: 081372822826
F. Nama Sekolah	: SDN 4 Bandung
G. Alamat Sekolah	: Jl. Diponegoro No.13
H. No.Telp Sekolah	: 084515647835
I. No. Kartu Pelajar	: 493431

Gambar 2. Potongan Contoh Dokumen Digital Formulir

Serta 5 dokumen digital berbentuk artikel yang memiliki kondisi memiliki huruf standar (sesuai kaidah penulisan Bahasa Indonesia yang benar) dan 5 dokumen artikel dengan huruf kapital secara penuh. Dokumen digital tersebut memiliki format *image* (png) sebagaimana terdapat pada Gambar 3 dan Gambar 4.



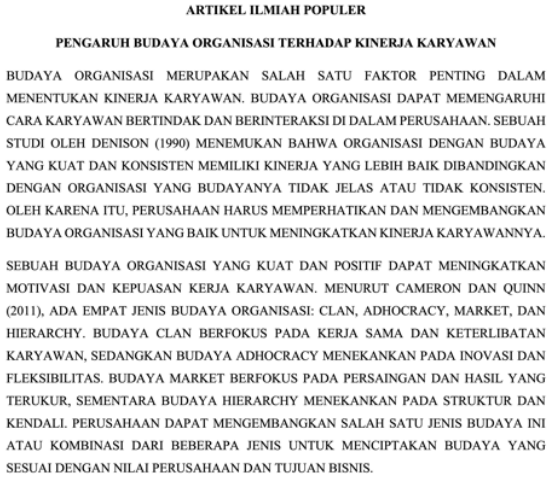
ARTIKEL ILMIAH POPULER

Pengaruh Budaya Organisasi terhadap Kinerja Karyawan

Budaya organisasi merupakan salah satu faktor penting dalam menentukan kinerja karyawan. Budaya organisasi dapat memengaruhi cara karyawan bertindak dan berinteraksi di dalam perusahaan. Sebuah studi oleh Denison (1990) menemukan bahwa organisasi dengan budaya yang kuat dan konsisten memiliki kinerja yang lebih baik dibandingkan dengan organisasi yang budayanya tidak jelas atau tidak konsisten. Oleh karena itu, perusahaan harus memperhatikan dan mengembangkan budaya organisasi yang baik untuk meningkatkan kinerja karyawannya.

Sebuah budaya organisasi yang kuat dan positif dapat meningkatkan motivasi dan kepuasan kerja karyawan. Menurut Cameron dan Quinn (2011), ada empat jenis budaya organisasi: clan, adhocracy, market, dan hierarchy. Budaya clan berfokus pada kerja sama dan keterlibatan karyawan, sedangkan budaya adhocracy menekankan pada inovasi dan fleksibilitas. Budaya market berfokus pada persaingan dan hasil yang terukur, sementara budaya hierarchy menekankan pada struktur dan kendali. Perusahaan dapat mengembangkan salah satu jenis budaya ini atau kombinasi dari beberapa jenis untuk menciptakan budaya yang sesuai dengan nilai perusahaan dan tujuan bisnis.

Gambar 3. Potongan Contoh Dokumen Digital Artikel dengan Huruf Sesuai Kaidah Penulisan



ARTIKEL ILMIAH POPULER

PENGARUH BUDAYA ORGANISASI TERHADAP KINERJA KARYAWAN

BUDAYA ORGANISASI MERUPAKAN SALAH SATU FAKTOR PENTING DALAM MENENTUKAN KINERJA KARYAWAN. BUDAYA ORGANISASI DAPAT MEMENGARUHI CARA KARYAWAN BERTINDAK DAN BERINTERAKSI DI DALAM PERUSAHAAN. SEBUAH STUDI OLEH DENISON (1990) MENEMUKAN BAHWA ORGANISASI DENGAN BUDAYA YANG KUAT DAN KONSISTEN MEMILIKI KINERJA YANG LEBIH BAIK DIBANDINGKAN DENGAN ORGANISASI YANG BUDAYANYA TIDAK JELAS ATAU TIDAK KONSISTEN. OLEH KARENA ITU, PERUSAHAAN HARUS MEMPERHATIKAN DAN MENGEMBANGKAN BUDAYA ORGANISASI YANG BAIK UNTUK MENINGKATKAN KINERJA KARYAWANNYA.

SEBUAH BUDAYA ORGANISASI YANG KUAT DAN POSITIF DAPAT MENINGKATKAN MOTIVASI DAN KEPUASAN KERJA KARYAWAN. MENURUT CAMERON DAN QUINN (2011), ADA EMPAT JENIS BUDAYA ORGANISASI: CLAN, ADHOCRACY, MARKET, DAN HIERARCHY. BUDAYA CLAN BERFOKUS PADA KERJA SAMA DAN KETERLIBATAN KARYAWAN, SEDANGKAN BUDAYA ADHOCRACY MENEKANKAN PADA INOVASI DAN FLEKSIBILITAS. BUDAYA MARKET BERFOKUS PADA PERSAINGAN DAN HASIL YANG TERUKUR, SEMENTARA BUDAYA HIERARCHY MENEKANKAN PADA STRUKTUR DAN KENDALI. PERUSAHAAN DAPAT MENGEMBANGKAN SALAH SATU JENIS BUDAYA INI ATAU KOMBINASI DARI BEBERAPA JENIS UNTUK MENCIPTAKAN BUDAYA YANG SESUAI DENGAN NILAI PERUSAHAAN DAN TUJUAN BISNIS.

Gambar 4. Potongan Contoh Dokumen Digital Artikel dengan Huruf Kapital

3. Implementasi OCR

Setelah dataset dipersiapkan, langkah selanjutnya yaitu menerapkan kedua *library* OCR (*EasyOCR* dan Tesseract) untuk proses ekstraksi teks dengan menggunakan aplikasi *Visual Studio Code*. Hasil data yang dikumpulkan meliputi waktu pemrosesan OCR, *Word Error Rate* (WER) untuk mengukur kesalahan pada tingkat kata, *Character Error Rate* (CER) untuk mengukur kesalahan pada tingkat karakter, dan

akurasi kata secara keseluruhan pada dataset dokumen (Marshanda et al., 2024).

4. Analisis akurasi pembacaan OCR

Langkah terakhir adalah proses evaluasi untuk mengukur seberapa baik sistem OCR dapat mengenali dan menyalin teks dari gambar atau dokumen ke dalam bentuk teks digital. Dalam konteks ini, terdapat beberapa metrik utama yang digunakan untuk mengevaluasi performa OCR, yaitu WER, CER, dan Akurasi OCR secara keseluruhan. Masing-masing metrik memiliki pendekatan dan kegunaan yang berbeda, tergantung pada tingkat granularitas dan tujuan evaluasi.

a. *Word Error Rate* (WER)

WER dihitung dengan membandingkan jumlah kata yang dihasilkan oleh *library* OCR dengan jumlah kata yang benar dalam teks asli. Sebuah penelitian menunjukkan bahwa penggunaan arsitektur yang lebih canggih, seperti *Long Short-Term Memory* (LSTM) dalam OCR, dapat meningkatkan akurasi WER secara signifikan, terutama dalam pengenalan teks yang kompleks (Setyadi & Susetyo, 2023). Untuk menghitung WER menggunakan persamaan sebagai berikut:

$$WER(\%) = \frac{S+D+I}{N} \times 100\% \quad (1)$$

di mana
S adalah jumlah kata yang salah dikenali.
D adalah jumlah kata yang hilang.
I adalah jumlah kata yang salah ditambahkan oleh OCR.
N adalah total jumlah kata dalam *ground truth*.

Ground truth adalah kumpulan data yang telah dianotasi dengan benar dan digunakan sebagai acuan untuk menilai akurasi OCR. *Ground truth* berfungsi sebagai standar atau referensi yang memungkinkan perbandingan antara hasil keluaran OCR dengan teks yang sebenarnya. Selain itu, dalam penelitian yang dilakukan oleh Maurer et al., diungkapkan bahwa *ground truth* yang dihasilkan dapat digunakan untuk menguji berbagai *library* OCR, termasuk *Kraken* dan *Tesseract* (Maurer et al., 2023).

b. *Character Error Rate* (CER)

CER memberikan metrik yang lebih detail dengan menghitung kesalahan pada tingkat karakter. Ini sangat berguna dalam konteks di mana kesalahan kecil dapat memiliki dampak besar seperti dalam pengenalan karakter untuk dokumen resmi atau sertifikat. Penelitian menunjukkan bahwa *library* OCR yang menggunakan teknik pemrosesan lanjutan, seperti pengolahan citra dan segmentasi karakter, dapat mengurangi CER secara signifikan (Apriyanti & Widodo, 2016). Untuk menghitung CER dapat menggunakan persamaan sebagai berikut:

$$CER(\%) = \frac{S_{char}+D_{char}+I_{char}}{N_{char}} \times 100\% \quad (2)$$

di mana
S_char adalah jumlah karakter yang salah dikenali.
D_char adalah jumlah karakter yang hilang.
I_char adalah jumlah karakter yang salah ditambahkan oleh OCR.
N_char adalah total jumlah karakter dalam *ground truth*.

c. Akurasi OCR secara keseluruhan

Untuk menghitung tingkat akurasi dengan membandingkan jumlah karakter yang benar dalam hasil OCR (*output*) terhadap total karakter yang ada dalam teks referensi (*ground truth*). Tujuannya adalah untuk mengukur seberapa akurat sistem OCR dalam mengenali dan menyalin teks dari gambar atau dokumen (Holila, 2024). Untuk menghitung akurasi menggunakan persamaan sebagai berikut:

$$\text{Akurasi}\ (\%) = \frac{Kata\ benar}{Total\ Kata} \times 100\% \quad (3)$$

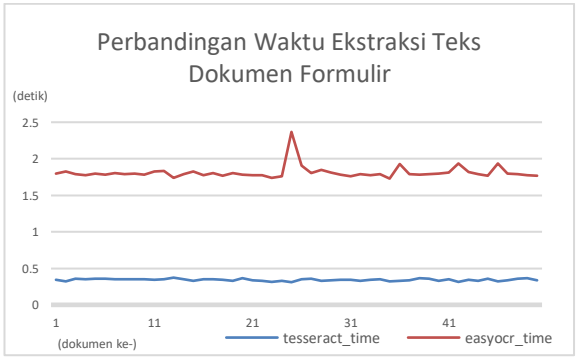
3. Hasil dan Pembahasan

A. Kinerja OCR pada dokumen formulir

Pada dataset pertama berisi dokumen formulir (50 gambar formulir), hasil pengujian menunjukkan perbedaan kinerja yang mencolok antara *Tesseract* dan *EasyOCR*, baik dari segi kecepatan maupun akurasi.

1) Waktu pemrosesan OCR

Setiap *library* menunjukkan pola waktu pemrosesan yang berbeda, memberikan gambaran mengenai efisiensi masing-masing dalam menangani dokumen dengan variasi struktur dan kualitas. Hal ini seperti ditunjukkan pada Gambar 5.



Gambar 5. Diagram Perbandingan Waktu Ekstraksi Teks Tesseract dan EasyOCR

Dari sisi waktu pemrosesan, *Tesseract* secara konsisten lebih cepat memproses gambar dibanding *EasyOCR*. *Tesseract* memiliki performa yang lebih stabil, dengan waktu pemrosesan yang relatif konstan pada rentang 0,31–0,36 detik per dokumen, dengan rata-rata 0,34 detik. Sebaliknya, *EasyOCR*

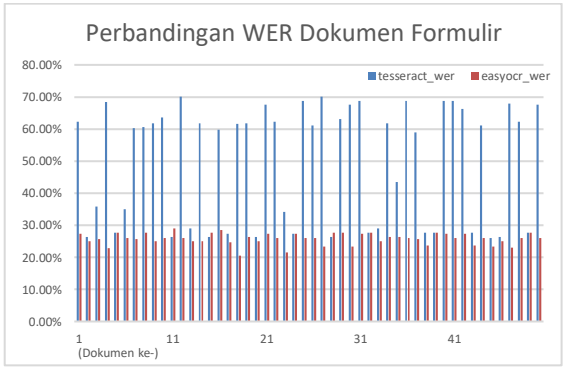
menunjukkan performa yang lebih bervariasi, dengan rentang waktu pemrosesan 1,73–2,36 detik per dokumen, dengan waktu rata-rata 1,81 detik

Dengan kata lain, *EasyOCR* membutuhkan waktu sekitar 5 kali lebih lama untuk mengekstrak teks dari gambar dibandingkan *Tesseract*. Hal ini menunjukkan efisiensi waktu *Tesseract* sehingga dapat diterapkan *real-time* atau untuk pemrosesan dokumen dalam jumlah besar, meskipun perlu diimbangi dengan pertimbangan akurasi hasil OCR (Patience et al., 2024).

2) Analisa pembacaan OCR

a. Perbandingan WER

WER yang dihasilkan oleh ekstraksi teks menggunakan *Tesseract* dan *EasyOCR* ditunjukkan pada Gambar 6.



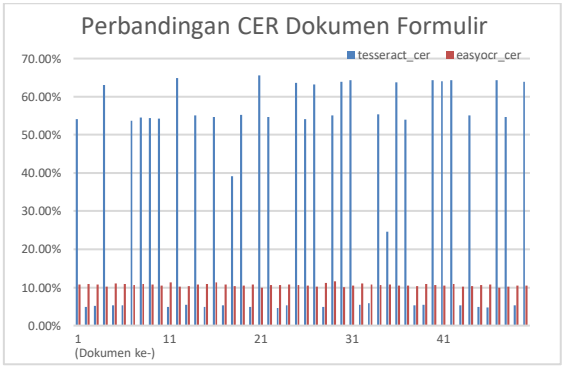
Gambar 6. Diagram Perbandingan WER Ekstraksi Teks Tesseract dan EasyOCR

Berdasarkan Gambar 6, *EasyOCR* menunjukkan WER yang lebih rendah dibandingkan dengan *Tesseract*. Rata-rata WER *EasyOCR* berada di angka 25,78%, sedangkan *Tesseract* memiliki rata-rata WER yang jauh lebih tinggi yaitu hingga 49,69%. Perbedaan ini mengindikasikan bahwa *EasyOCR* lebih unggul dalam mengenali kata dengan benar dan lebih stabil dalam berbagai kondisi dokumen dibandingkan dengan *Tesseract*, yang lebih rentan terhadap kesalahan segmentasi dan misidentifikasi kata.

Dari distribusi data, WER *Tesseract* sangat bervariasi dengan beberapa dokumen memiliki tingkat kesalahan setinggi 70,13%, sementara beberapa lainnya hanya sekitar 25,97%. Ini menunjukkan bahwa *Tesseract* tidak memiliki kestabilan dalam pengenalan kata, karena performanya sangat bergantung pada kualitas dokumen, seperti keterbacaan teks, resolusi gambar, serta adanya *noise* atau gangguan lain. Sebaliknya, *EasyOCR* lebih unggul namun dari data menunjukkan pola kesalahan yang lebih seragam, dengan sebagian besar dokumen memiliki WER dalam rentang 23,38% hingga 28,95%, yang berarti sistem ini mampu mempertahankan akurasi yang lebih baik bahkan pada dokumen dengan format yang lebih kompleks.

b. Perbandingan CER

CER yang dihasilkan oleh ekstraksi teks menggunakan *Tesseract* dan *EasyOCR* ditunjukkan pada Gambar 7.



Gambar 7. Diagram Perbandingan CER Ekstraksi Teks Tesseract dan EasyOCR

Dari Gambar 7, CER menunjukkan bahwa *EasyOCR* memiliki tingkat kesalahan karakter yang jauh lebih rendah dibandingkan *Tesseract* dengan CER rata-rata *EasyOCR* sekitar 10,64%. Sedangkan CER rata-rata *Tesseract* sekitar 36,40%. Ini menunjukkan bahwa *EasyOCR* menghasilkan lebih sedikit kesalahan dalam mengenali karakter secara individual, sehingga lebih akurat dibandingkan *Tesseract* dalam berbagai kondisi dokumen.

Pada sebagian besar dokumen uji, *EasyOCR* menunjukkan CER yang stabil dalam kisaran 9,89% hingga 11,61%, yang menandakan bahwa metode ini dapat mengenali lebih banyak karakter dengan benar dan lebih tahan terhadap variasi dalam struktur teks. Sebagai perbandingan, CER *Tesseract* sangat bervariasi, dengan beberapa dokumen memiliki tingkat kesalahan setinggi 65,59%, sementara yang lain turun hingga sekitar 4,68%.

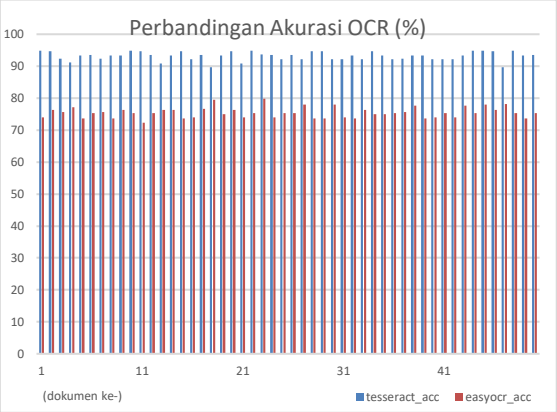
Pola ini menunjukkan bahwa *Tesseract* lebih sensitif terhadap kualitas dokumen. Pada dokumen dengan struktur yang lebih jelas dan *font* standar, CER *Tesseract* bisa turun hingga 4,68%–5,82%, tetapi dalam banyak kasus, terutama dengan tata letak yang lebih kompleks, CER dapat melonjak jauh di atas 60%, menandakan tingkat kesalahan yang signifikan dalam segmentasi karakter. Sebaliknya, *EasyOCR* tetap lebih stabil, bahkan dalam kondisi dokumen yang lebih kompleks, dengan CER tidak melebihi 10,64%, yang berarti bahwa setidaknya hampir 90% karakter dikenali dengan benar.

c. Perbandingan akurasi OCR secara keseluruhan

Hasil dari pengukuran akurasi kedua *library* OCR dalam mengenali teks secara keseluruhan ditunjukkan pada Gambar 8.

Berdasarkan Gambar 8, akurasi OCR antara *Tesseract* dan *EasyOCR* menunjukkan perbedaan yang signifikan dalam cara kedua metode ini mengenali teks dalam dokumen gambar. Secara rata-rata, *Tesseract* memiliki akurasi sebesar 93,08%, sedangkan *EasyOCR* memiliki rata-rata akurasi

sebesar 75,42%. Selisih yang cukup besar antara kedua metode ini menunjukkan bahwa *Tesseract* secara keseluruhan mampu mengenali lebih banyak kata yang benar dibandingkan *EasyOCR*. Namun, meskipun angka ini menunjukkan keunggulan *Tesseract* dalam hal kuantitas kata yang dikenali, tidak serta-merta berarti bahwa hasil yang diberikan lebih akurat dalam struktur dan konteks teks yang sebenarnya.



Gambar 8. Diagram Perbandingan Akurasi Ekstraksi Teks Tesseract dan EasyOCR

Dari pola distribusi akurasi, *Tesseract* menunjukkan konsistensi dengan hampir seluruh sampel berada dalam rentang 90% hingga 95%, dengan hanya beberapa yang turun sedikit di bawah 90%. Sebaliknya, *EasyOCR* memiliki fluktuasi yang lebih besar dengan akurasi berkisar antara 72% hingga 79%. Hal ini mengindikasikan bahwa *Tesseract* lebih stabil dalam mengenali teks dengan tingkat keberhasilan yang tinggi dalam setiap sampel dokumen, sementara *EasyOCR* menunjukkan variasi yang lebih luas yang mungkin disebabkan oleh faktor kompleksitas tata letak dokumen atau kualitas teks dalam gambar. Namun, ada beberapa sampel dalam data di mana *EasyOCR* menunjukkan akurasi lebih tinggi dibandingkan rata-rata, menunjukkan bahwa dalam kondisi tertentu, metode berbasis *deep learning* ini dapat bekerja dengan sangat baik (Salehudin et al., 2023).

Untuk menganalisis kinerja *library* OCR *Tesseract* dan *EasyOCR* lebih mendalam dapat melihat perbandingan waktu pemrosesan, WER, CER, dan akurasi sebagaimana pada Tabel 1.

Tabel 1. Perbandingan Waktu Pemrosesan, WER, CER dan Akurasi

Metrik	Tesseract	EasyOCR
Waktu Pemrosesan Rata-rata (detik)	0,34 detik	1,81 detik
Word Error Rate (WER) Rata-rata	49,69%	25,78%
Character Error Rate (CER) Rata-rata	36,40%	10,64%
Akurasi keseluruhan	93,21%	75,50%

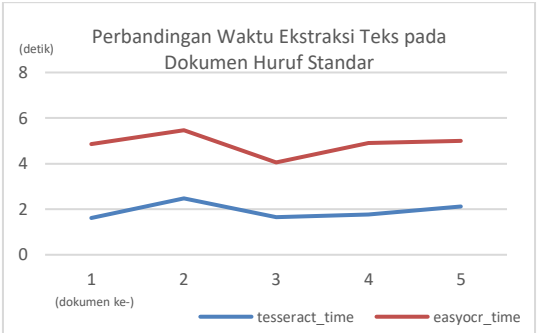
Berdasarkan analisis data akurasi antara *Tesseract* dan *EasyOCR* pada Tabel 1, terlihat bahwa *Tesseract* memiliki tingkat akurasi rata-rata yang lebih tinggi dibandingkan *EasyOCR* yaitu sekitar 93,21% dibandingkan dengan 75,50%. Hal ini menunjukkan bahwa *Tesseract* lebih unggul dalam mengenali kata-kata dalam dokumen gambar. Namun jika dibandingkan dengan metrik WER dan CER, terlihat adanya anomali yang menunjukkan bahwa tingkat kesalahan *Tesseract* dalam mengenali teks lebih tinggi dibandingkan *EasyOCR*. WER rata-rata *Tesseract* mencapai 49,69%, yang berarti hampir setengah dari seluruh kata dalam dokumen dikenali secara keliru, sedangkan *EasyOCR* hanya memiliki rata-rata WER sebesar 25,78% yang berarti lebih sedikit kesalahan dalam mengenali kata. Begitu pula dengan CER, di mana *Tesseract* mencatatkan tingkat kesalahan karakter sebesar 36,40%, jauh lebih tinggi dibandingkan EasyOCR yang hanya 10,64%. Hal ini mengindikasikan bahwa meskipun *Tesseract* mampu mengenali lebih banyak kata dalam dokumen walaupun kualitas pengenalan tersebut tidak selalu tepat dan seringkali terjadi kesalahan dalam struktur atau format teks yang dihasilkan dari dokumen formulir.

B. Kinerja OCR pada dokumen artikel

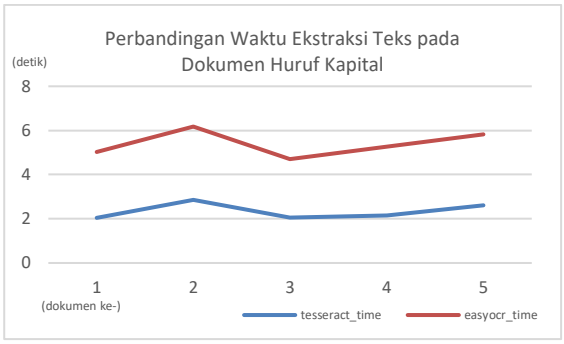
Dataset kedua terdiri atas 10 dokumen artikel dengan kondisi 5 dokumen menggunakan huruf sesuai kaidah penulisan Bahasa Indonesia dan 5 dokumen lainnya menggunakan huruf kapital secara keseluruhan. Konfigurasi ini sengaja dirancang untuk tujuan perbandingan dengan dokumen formulir, mengingat terdapat perbedaan mendasar dalam karakteristik format penulisan antara kedua jenis dokumen tersebut. Dokumen artikel ini memiliki variasi struktur teks yang lebih kompleks dan pola penggunaan huruf yang heterogen jika dibandingkan dengan dokumen formulir yang bersifat terstruktur dan seragam.

1) Waktu pemrosesan OCR

Hasil waktu permrosesan OCR pada dokumen artikel dengan kondisi huruf standar (sesuai kaidah penulisan Bahasa Indonesia) dan huruf kapital secara keseluruhan ditunjukan pada Gambar 9 dan Gambar 10.



Gambar 9. Diagram Perbandingan Waktu Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Standar



Gambar 10. Diagram Perbandingan Waktu Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Kapital

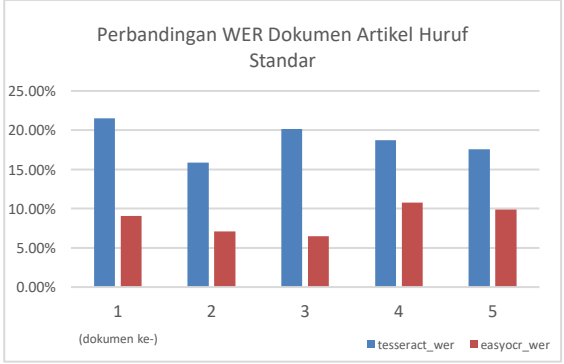
Pada dokumen dengan huruf standar seperti terlihat di Gambar 9, bahwa waktu pemrosesan rata-rata untuk *Tesseract* lebih cepat dibandingkan *EasyOCR*. *Tesseract* mencatat rata-rata waktu sekitar 1,93 detik per dokumen, sementara *EasyOCR* membutuhkan waktu lebih lama, yakni sekitar 4,85 detik. Selisih waktu yang cukup signifikan ini mengindikasikan bahwa dari aspek kecepatan, *Tesseract* memiliki keunggulan lebih efisien dibandingkan *EasyOCR*.

Sementara itu, pada dokumen dengan huruf kapital secara keseluruhan seperti pada Gambar 10, waktu pemrosesan mengalami peningkatan secara signifikan dibandingkan dengan dokumen yang menggunakan huruf standar, baik pada *Tesseract* maupun *EasyOCR*. *Tesseract* mencatat waktu rata-rata sebesar 2,34 detik, sedangkan *EasyOCR* meningkat tajam menjadi sekitar 5,40 detik. Ini menunjukkan bahwa dokumen dengan huruf kapital seluruhnya memang cenderung lebih sulit diproses, menyebabkan peningkatan waktu pemrosesan pada kedua OCR, dengan *EasyOCR* sekali lagi membutuhkan waktu jauh lebih lama dibandingkan *Tesseract*.

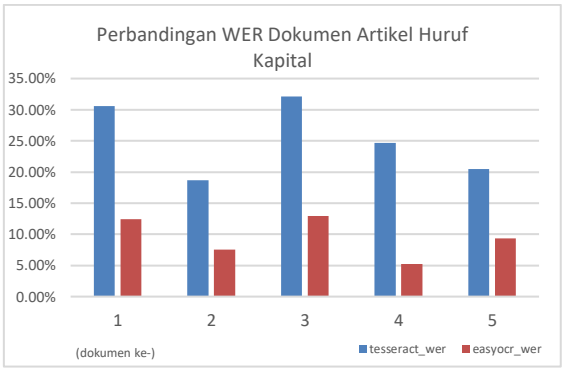
2) Analisa Pembacaan OCR

a. Perbandingan WER

Hasil WER ekstraksi OCR pada dokumen artikel dengan kondisi huruf standar (sesuai kaidah penulisan Bahasa Indonesia) dan huruf kapital secara keseluruhan ditunjukkan pada Gambar 11 dan Gambar 12.



Gambar 11. Diagram Perbandingan WER Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Standar



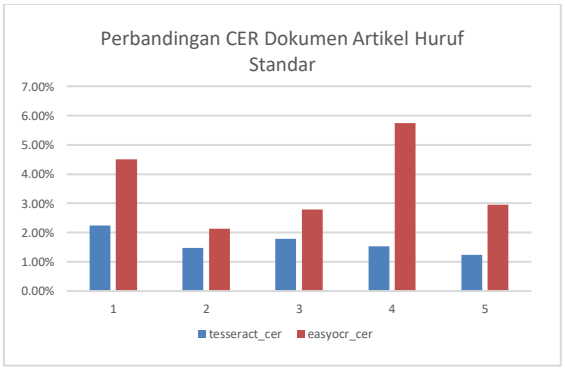
Gambar 12. Diagram Perbandingan WER Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Kapital

Pada Gambar 11, *EasyOCR* mencatat rata-rata WER sebesar 8,65%, lebih rendah dibandingkan dengan *Tesseract* yang mencapai 18,76%. Ini menunjukkan bahwa secara umum *EasyOCR* memiliki kemampuan lebih baik dalam mengenali kata dibanding *Tesseract* saat mengekstrak teks yang ditulis sesuai dengan kaidah huruf sesuai tata bahasa. Hasil WER per dokumen terbaik ditunjukan pada *EasyOCR* 7,11%, dan *Tesseract* 15,85%.

Pada kondisi huruf kapital seluruhnya seperti terlihat pada Gambar 12, rata-rata WER *EasyOCR* meningkat menjadi 9,51% dibandingkan Tesseract yang mengalami kenaikan signifikan menjadi 25,29%. Hal ini menunjukkan bahwa kedua *library* OCR lebih sulit mengenali teks dengan kapital secara keseluruhan, tetapi *EasyOCR* tetap mempertahankan performa yang lebih baik dibandingkan *Tesseract*. Secara individual, performa terbaik *EasyOCR* dalam skenario huruf kapital terdapat pada 5,21%, sementara hasil terbaik *Tesseract* terdapat pada 18,70%.

b. Perbandingan CER

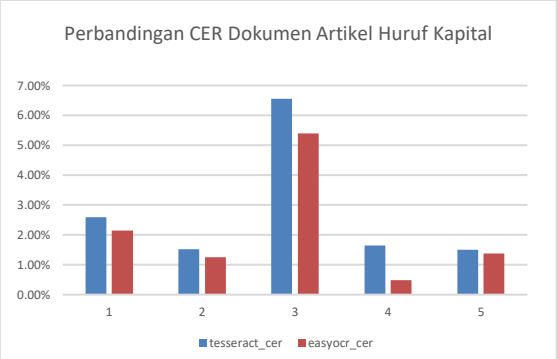
Hasil CER ekstraksi OCR pada dokumen artikel dengan kondisi huruf standar (sesuai kaidah penulisan Bahasa Indonesia) dan huruf kapital secara keseluruhan ditunjukkan pada Gambar 13 dan Gambar 14.



Gambar 13. Diagram Perbandingan CER Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Standar

Berdasarkan Gambar 13, uji kesalahan tingkat karakter pada *Tesseract* menunjukkan rata-rata CER sebesar 1,65%, lebih baik dibandingkan *EasyOCR*

yang mencatatkan nilai sebesar 3,62%. Hal ini memperlihatkan bahwa meskipun *Tesseract* memiliki tingkat kesalahan lebih tinggi dalam mengenali kata secara utuh, namun secara detail karakter, *Tesseract* lebih presisi dibanding *EasyOCR* pada dokumen huruf standar dengan CER terbaik *Tesseract* adalah pada 1,24%, sementara *EasyOCR* memiliki CER terbaik 2,13%.

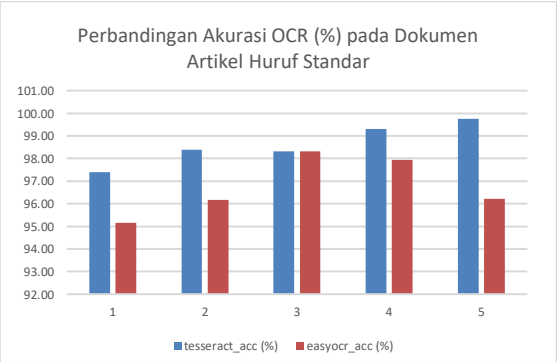


Gambar 14. Diagram Perbandingan CER Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Kapital

Dalam perbandingan CER pada dokumen huruf kapital penuh pada Gambar 14, justru *EasyOCR* unggul dengan nilai CER rata-rata sebesar 2,13% dan lebih rendah dibandingkan *Tesseract* yang mencatatkan CER sebesar 2,76%. Ini menunjukkan *EasyOCR* lebih baik dalam presisi karakter dibanding *Tesseract* ketika huruf menggunakan kapital seluruhnya. Untuk detail per dokumen, *EasyOCR* memiliki nilai CER terbaik mencapai 0,48%, sementara *Tesseract* mencatatkan hasil terbaik 1,53%.

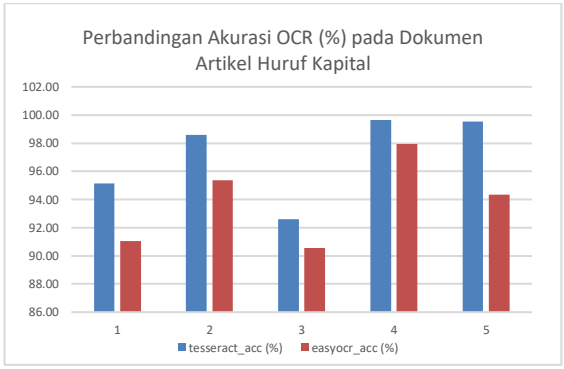
c. Perbandingan akurasi OCR secara keseluruhan

Hasil akurasi ekstraksi OCR pada dokumen artikel dengan kondisi huruf standar (sesuai kaidah penulisan Bahasa Indonesia) dan huruf kapital secara keseluruhan ditunjukan pada Gambar 13 dan Gambar 14.



Gambar 15. Diagram Perbandingan Akurasi Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Standar

Dari Gambar 15, rata-rata akurasi secara keseluruhan menunjukkan bahwa *Tesseract* sedikit lebih unggul (98,63%) dibandingkan *EasyOCR* (96,76%) pada kasus dokumen dengan penulisan sesuai kaidah. Ini menyiratkan bahwa secara umum *Tesseract* lebih handal dalam menghasilkan teks dengan kesalahan minimal dibanding *EasyOCR* dalam skenario ini. Akurasi terbaik *Tesseract* dicapai pada nilai 99,76%, sedangkan *EasyOCR* menunjukkan akurasi terbaiknya pada 98,32%.



Gambar 16. Diagram Perbandingan Akurasi Ekstraksi Teks Tesseract dan EasyOCR pada Dokumen Artikel Huruf Kapital

Sedangkan dalam skenario huruf kapital secara keseluruhan seperti Gambar 16, menunjukkan bahwa *Tesseract* kembali mencatatkan performa lebih baik dengan rata-rata akurasi sebesar 97,10%, dibandingkan dengan *EasyOCR* yang mencatatkan nilai akurasi sebesar 93,85%. Meskipun terjadi penurunan dibandingkan skenario pertama, *Tesseract* tetap lebih unggul secara umum dalam mempertahankan akurasi tinggi. Dalam detail tiap dokumen, *Tesseract* memiliki akurasi terbaik pada Artikel 4 sebesar 99,66%, sedangkan *EasyOCR* mencapai akurasi tertinggi pada dokumen yang sama, yaitu Artikel 4 sebesar 97,95%.

Untuk menganalisis kinerja *library* OCR Tesseract dan *EasyOCR* lebih mendalam berdasarkan dapat melihat perbandingan waktu pemrosesan, WER, CER, dan akurasi sebagaimana pada Tabel 2.

Tabel 2. Perbandingan Waktu Pemrosesan, WER, CER dan Akurasi Dokumen Artikel

Metrik	**Dokumen Huruf Standar**		**Dokumen Huruf Kapital**	
	Tesseract	**EasyOCR**	**Tesseract**	**EasyOCR**
Waktu Pemrosesan Rata-rata (detik)	1,93 detik	4,85 detik	2,34 detik	5,40 detik
Word Error Rate (WER) Rata-rata	18,76%	8,65%	25,29%	9,51%
Character Error Rate (CER) Rata-rata	1,65%	3,62%	2,76%	2,13%
Akurasi keseluruhan	98,63%	96,76%	97,10%	93,85%

Berdasarkan Tabel 2, perbandingan antara metode OCR *Tesseract* dan *EasyOCR* menunjukkan hubungan yang jelas antara kecepatan pemrosesan dengan tingkat akurasi pengenalan teks. Secara umum, *Tesseract* memiliki kecepatan pemrosesan

rata-rata yang jauh lebih baik dibandingkan *EasyOCR*, baik untuk dokumen dengan huruf standar (1,93 detik dibandingkan 4,85 detik) maupun dokumen dengan huruf kapital (2,34 detik dibandingkan 5,40 detik). Namun, *Tesseract* memiliki kelemahan dalam aspek WER yang lebih tinggi dibandingkan *EasyOCR*, terutama pada dokumen dengan huruf kapital (25,29% dibandingkan 9,51%) serta dokumen huruf standar (18,76% dibandingkan 8,65%). Hal ini menunjukkan bahwa *EasyOCR*, meskipun lebih lambat dalam pemrosesan, mampu menghasilkan akurasi kata yang jauh lebih baik dibandingkan *Tesseract*.

Selain itu, performa kedua *library* OCR ini ternyata berbeda signifikan tergantung pada format teks dokumen. Pada dokumen dengan huruf kapital, kedua OCR mengalami penurunan kinerja yang terlihat dari meningkatnya waktu pemrosesan, WER, serta penurunan tingkat akurasi keseluruhan dibanding dokumen standar. Penurunan ini lebih mencolok pada *Tesseract* yang mengalami peningkatan WER dari 18,76% (huruf standar) menjadi 25,29% (huruf kapital), sementara *EasyOCR* mengalami peningkatan yang relatif kecil dari 8,65% menjadi 9,51%. Kondisi ini menunjukkan bahwa *Tesseract* cenderung lebih sensitif terhadap format teks, sementara *EasyOCR* lebih konsisten dan stabil di berbagai jenis format teks.

Dilihat dari hubungan antara CER dengan WER, *Tesseract* secara mengejutkan memiliki nilai CER yang lebih rendah untuk dokumen standar (1,65%) dibanding *EasyOCR* (3,62%), yang seharusnya mencerminkan akurasi karakter lebih tinggi. Namun *Tesseract* memiliki WER yang jauh lebih tinggi (18,76%) dibanding *EasyOCR* (8,65%) menunjukan bahwa kesalahan yang dibuat *Tesseract* seringkali terjadi pada level kata secara keseluruhan, bukan hanya pada tingkat karakter. Sebaliknya, *EasyOCR* memiliki kemampuan mengenali keseluruhan kata yang lebih baik meskipun kesalahan di level karakter sedikit lebih tinggi. Kondisi serupa juga terjadi pada dokumen huruf kapital, dimana *EasyOCR* memiliki CER lebih rendah (2,13%) dibandingkan *Tesseract* (2,76%), sekaligus memperkuat bahwa *EasyOCR* lebih stabil secara keseluruhan dalam mengenali kata-kata secara utuh. Dengan demikian, jika prioritas utama adalah kecepatan pemrosesan dan akurasi karakter yang tinggi, *Tesseract* lebih cocok digunakan dokumen huruf standar dengan akurasi keseluruhan mencapai 98,63%. Namun, jika tujuan utama penggunaannya adalah mendapatkan tingkat akurasi kata yang tinggi serta stabil dalam berbagai jenis format teks, maka *EasyOCR* dengan WER yang lebih rendah (8,65%-9,51%) lebih baik digunakan meskipun waktu pemrosesannya relatif lebih lama dibandingkan *Tesseract*.

4. Kesimpulan

Berdasarkan analisis kinerja *library* OCR pada dokumen formulir dan artikel, *Tesseract* menunjukkan keunggulan dalam kecepatan pemrosesan, dengan waktu rata-rata 0,34 detik per dokumen formulir, sementara *EasyOCR* membutuhkan 1,81 detik. Keunggulan ini juga terlihat pada dokumen artikel, di mana *Tesseract* mencatat waktu pemrosesan rata-rata 1,93 detik (huruf standar) dan 2,34 detik (huruf kapital), lebih cepat dibanding *EasyOCR* dengan waktu 4,85 detik dan 5,40 detik.

Namun, dalam hal akurasi pengenalan kata (WER), *EasyOCR* secara jelas lebih unggul. Pada dokumen formulir, *EasyOCR* memiliki WER sebesar 25,78% lebih rendah dibanding *Tesseract* yang mencapai 49,69%. Kondisi serupa juga terjadi pada dokumen artikel, dimana *EasyOCR* mencatatkan WER sebesar 8,65% (huruf standar) dan 9,51% (huruf kapital), lebih rendah dibanding *Tesseract* yang masing-masing mencapai 18,76% dan 25,29%.

Dengan analisis mendalam, meskipun *Tesseract* unggul dalam kecepatan dan memiliki akurasi karakter tinggi pada artikel huruf standar namun *EasyOCR* menunjukkan kestabilan dan konsistensi lebih baik dalam mengenali kata secara akurat, terutama pada dokumen dengan kompleksitas tinggi (formulir) dan penggunaan huruf kapital secara penuh (artikel). Sehingga untuk ekstraksi OCR yang membutuhkan akurasi tinggi pada pengenalan kata khususnya pada dokumen kompleks, lebih tepat menggunakan *EasyOCR*. Sebaliknya, untuk kebutuhan pemrosesan cepat dengan kondisi dokumen yang lebih sederhana dan jelas, *Tesseract* merupakan pilihan yang lebih efisien.

Daftar Pustaka:

Al amin, I. H., & Aprilino, A. (2022). IMPLEMENTASI ALGORITMA YOLO DAN TESSERACT OCR PADA SISTEM DETEKSI PLAT NOMOR OTOMATIS. *Jurnal Teknoinfo*, *16*(1), 54. https://doi.org/10.33365/jti.v16i1.1522

Marshanda, Harijanto, B., & Rahmad, C. (2024). Implementasi Optical Character Recognition (OCR) untuk Meningkatkan Akurasi dan Kecepatan Input Data di Posyandu. *Jurnal Informatika Polinema*, *11*(1), 45–50. https://doi.org/10.33795/jip.v11i1.6025

Apriyanti, K., & Widodo, T. (2016). Implementasi Optical Character Recognition Berbasis Backpropagation untuk Text to Speech Perangkat Android. *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, *6*(1), 13. https://doi.org/10.22146/ijeis.10767

Asroni, A., Indrawan, G., & Erawati Dewi, L. J. (2023). Implementasi Hirarki Dataset Dalam Membangun Model Language Aksara Bali Menggunakan Framework Tesseract OCR. *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, *6*(1), 20–28.

https://doi.org/10.31598/jurnalresistor.v6i1.1345

Banu, K., Andreas, D., Anggoro, W., & Setiawan, A. (2023). OCR: Masa Depan Pengenalan Karakter Optik dan Dampaknya pada Kehidupan Modern. *Jurnal Teknologi Informasi*, *9*(2), 147–156. https://doi.org/10.52643/jti.v9i2.3798

Hamdi, A., Chan, Y. K., & Koo, V. C. (2021). A New Image Enhancement and Super Resolution technique for license plate recognition. *Heliyon*, *7*(11), e08341. https://doi.org/10.1016/j.heliyon.2021.e08341

Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science*, *5*(1), 861–882. https://doi.org/10.1007/s42001-021-00149-1

Holila, A. R. P. S. A. P. L. J. I. (2024). *Introduction National Identification Number and Name on Id Card Using Ocr (Optical Character Recognition) Method*. https://doi.org/10.52436/1.jutif.2024.5.4.2242

Maurer, Y., Schneider, P., & Marschall, R. (2023). Nautilus. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *33*(1), 1–19. https://doi.org/10.53377/lq.13330

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2022). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, *54*(6), 1–37. https://doi.org/10.1145/3453476

Patience, O. O., Amaechi, E. M., George, O., & Isaac, O. N. (2024). Enhanced Text Recognition in Images Using Tesseract OCR within the Laravel Framework. *Asian Journal of Research in Computer Science*, *17*(9), 58–69. https://doi.org/10.9734/ajrcos/2024/v17i9499

Raswaty, H. S., & Nuryuliani, N. (2021). Implementation of Optical Character Recognition and Voice Recognition of House of Words (How) Dictionary Application on Android Platform. *Engineering, MAthematics and Computer Science (EMACS) Journal*, *3*(3), 93–101. https://doi.org/10.21512/emacsjournal.v3i3.7418

Salehudin, M. A. M., Basah, S. N., Yazid, H., Basaruddin, K. S., Safar, M. J. A., Som, M. H. M., & Sidek, K. A. (2023). Analysis of Optical Character Recognition using EasyOCR under Image Degradation. *Journal of Physics: Conference Series*, *2641*(1), 012001. https://doi.org/10.1088/1742-6596/2641/1/012001

Setyadi, A. F. I., & Susetyo, Y. A. (2023). Implementasi Algoritma LSTM pada Aplikasi Optical Character Recognition Berbasis Website Menggunakan Tesseract OCR. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *6*(2), 63–71. https://doi.org/10.32493/jtsi.v6i2.29235

Sharma, A., Ansari, A. Z., Kakulavarapu, R., Stensen, M. H., Riegler, M. A., & Hammer, H. L. (2023). Predicting Cell Cleavage Timings from Time-Lapse Videos of Human Embryos. *Big Data and Cognitive Computing*, *7*(2), 91. https://doi.org/10.3390/bdcc7020091

Sporici, D., Cușnir, E., & Boiangiu, C.-A. (2020). Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. *Symmetry*, *12*(5), 715. https://doi.org/10.3390/sym120507

Halaman ini sengaja dikosongkan