

Dampak Prapemrosesan Citra dan Teks Terhadap Tingkat Akurasi OCR dan Skor ROUGE

Fajar Hamdani^{1*}, Esa Fauzi²

^{1,2} Fakultas Teknik, Universitas Widyatama, Bandung, Indonesia

Email: ¹fajar.hamdani@widyatama.ac.id, ²esa.fauzi@widyatama.ac.id

Abstrak

Dokumen digital berformat PDF yang memuat tulisan tangan atau hasil pindaian beresolusi rendah sering kali sulit untuk diekstraksi informasinya. Penelitian ini mengkaji dampak prapemrosesan citra dan teks dalam meningkatkan akurasi proses ekstraksi menggunakan teknologi *Optical Character Recognition* (OCR) serta peringkasan otomatis menggunakan *Natural Language Processing* (NLP). Fokus penelitian tidak diarahkan pada perancangan model OCR atau NLP baru, melainkan pada signifikansi tahapan prapemrosesan terhadap kualitas teks masukan pada sistem. Tahapan prapemrosesan citra yang diterapkan bersifat adaptif, meliputi pemotongan area teks, *deskew*, *resize* bikubik, konversi *grayscale*, *denoising*, *Contrast Limited Adaptive Histogram Equalization* (CLAHE), *Gaussian blur*, *adaptive threshold*, dan morfologi *opening*. Teks hasil OCR kemudian dibersihkan dan diperbaiki secara adaptif sebelum diringkas menggunakan model T5 bahasa Indonesia. Hasil pengujian terhadap 23 sampel dokumen menunjukkan bahwa intervensi prapemrosesan berhasil meningkatkan akurasi pengenalan karakter dari 65,31% menjadi 75,10%. Selain itu, pada tahap evaluasi peringkasan teks juga terjadi peningkatan skor ROUGE-1 dari 21,89% menjadi 32,24%, ROUGE-2 dari 8,69% menjadi 16,16%, dan ROUGE-L dari 16,93% menjadi 26,43%. Hasil pengujian ini membuktikan bahwa tahapan prapemrosesan citra dan teks sangat penting dalam memperbaiki kinerja pengenalan karakter dan peringkasan dokumen. Penelitian prapemrosesan adaptif ini menekankan bahwa tidak semua tahapan dilakukan, namun menyeleksi tahap prapemrosesan citra berdasarkan dampak terhadap peningkatan akurasi pengenalan karakter serta kenaikan ROUGE sebelum ringkasan akhir digunakan.

Kata Kunci: Prapemrosesan Citra; OCR; PaddleOCR; NLP; T5;

Abstract

Digital documents in PDF format containing handwritten or low-resolution scans are often difficult to extract information from. This study examines the impact of image and text preprocessing in improving the accuracy of the extraction process using Optical Character Recognition (OCR) technology and automatic summarization using Natural Language Processing (NLP). The focus of the study is not directed at designing a new OCR or NLP model, but rather on the significance of the preprocessing stage on the quality of the input text in the system. The image preprocessing stages applied are adaptive, including cropping the text area, *deskew*, *bicubic resizing*, *grayscale conversion*, *denoising*, *Contrast Limited Adaptive Histogram Equalization* (CLAHE), *Gaussian blur*, *adaptive threshold*, and *morphological opening*. The OCR text is then cleaned and improved adaptively before being summarized using the Indonesian T5 model. Test results on 23 document samples show that the preprocessing intervention successfully increased the accuracy of character recognition from 65.31% to 75.10%. In addition, in the text summarization evaluation stage, there was also an increase in the ROUGE-1 score from 21.89% to 32.24%, ROUGE-2 from 8.69% to 16.16%, and ROUGE-L from 16.93% to 26.43%. These test results prove that the image and text preprocessing stages are very important in improving the performance of character recognition and document summarization. This adaptive preprocessing research emphasizes that not all stages are performed, but rather selects the image preprocessing stages based on their impact on improving character recognition accuracy and increasing ROUGE scores before the final summary is utilized.

Keywords: Image Preprocessing; OCR; PaddleOCR; NLP; T5;

*Corresponding Author:

Fajar Hamdani, Universitas Widyatama, Indonesia

Email: fajar.hamdani@widyatama.ac.id

1.PENDAHULUAN

Perkembangan digitalisasi membuat pemrosesan dokumen semakin sering dilakukan secara otomatis, terutama pada pekerjaan administrasi, arsip, dan pengelolaan data. Salah satu teknologi yang banyak digunakan adalah *Optical Character Recognition* (OCR), yaitu teknologi untuk mengubah informasi visual pada dokumen seperti *Portable Document Format* (PDF) menjadi teks yang dapat diolah komputer. OCR sudah digunakan pada berbagai kebutuhan, mulai dari entri data, pelacakan dokumen, sampai integrasi dengan sistem informasi [26]. Namun, hasil OCR masih sangat dipengaruhi oleh kualitas dokumen. Dokumen yang berisi campuran teks cetak dan tulisan tangan biasanya lebih sulit dibaca, terutama jika citranya buram, miring, memiliki kontras rendah, atau terkena bayangan saat dipindai [22]. Kesalahan kecil pada tahap pengenalan karakter dapat berubah menjadi kesalahan kata, lalu ikut memengaruhi proses berikutnya ketika teks tersebut digunakan untuk peringkasan otomatis.

Sejumlah penelitian telah mencoba menurunkan kesalahan OCR melalui prapemrosesan citra. Dey et al. [8] membahas beberapa teknik seperti binarisasi, penipisan (*thinning*), dan pengurangan derau, serta menunjukkan bahwa keberhasilan OCR bergantung pada kesesuaian teknik prapemrosesan dengan kondisi dokumen. Mursari and Wibowo [13] juga menunjukkan bahwa pembersihan derau dan perataan piksel dapat membantu pengenalan tulisan tangan digital menggunakan Tesseract. Walaupun demikian, metode tersebut belum selalu berhasil pada tulisan tangan yang bentuk hurufnya sangat beragam. Talirongan et al. [22] menguji strategi yang memadukan *Warp Perspective*, *Gaussian Blurring*, dan binarisasi Otsu. Hasilnya meningkat pada teks cetak dan teks campuran, tetapi pengenalan tulisan tangan murni masih rendah karena bentuk hurufnya tidak seragam.

Beberapa penelitian lain mulai menggabungkan pengolahan citra dengan model *deep learning* agar sistem lebih tahan terhadap variasi bentuk huruf. Wu and Chang [26], misalnya, menggunakan beberapa tahap prapemrosesan seperti koreksi kemiringan, CLAHE, dan dekonvolusi Wiener sebelum teks diekstraksi. Bhaskoro et al. [5] meneliti interpretasi tulisan tangan bahasa Indonesia dengan memadukan pengolahan citra dan model CRNN. Selain perbaikan pada citra, koreksi setelah OCR juga mulai banyak digunakan. Almegbil et al. [2] memperkenalkan Raqim, yaitu pendekatan hibrida yang menggabungkan kamus kesalahan dan model bahasa generatif untuk memperbaiki teks hasil OCR. Dari berbagai studi tersebut, dapat terlihat bahwa kualitas data yang masuk (citra dan teks) mempengaruhi keberhasilan sistem.

Berdasarkan hasil kajian tersebut, prapemrosesan menjadi fokus utama dalam penelitian ini. Sebagian penelitian sebelumnya masih membahas prapemrosesan citra, OCR, dan NLP secara terpisah. Pada penelitian ini, ketiganya digabungkan dalam satu alur kerja, mulai dari perbaikan citra, pengenalan tulisan tangan dengan PaddleOCR, pembersihan teks, hingga peringkasan menggunakan model T5 bahasa Indonesia (*cahya/t5-base-indonesian-summarization-cased*). Tahapan prapemrosesan citra yang digunakan meliputi memotong area tulisan, *deskew*, *resize bicubic*, *grayscale*, *denoising*, CLAHE, *Gaussian blur*, *adaptive threshold*, dan morfologi open. Kemudian T5 dipilih karena mampu melakukan peringkasan abstraktif, yaitu menyusun kembali informasi menjadi kalimat ringkasan baru, bukan sekadar mengambil kalimat dari teks asli [15].

Berangkat dari permasalahan diatas, penelitian ini difokuskan untuk menjawab bagaimana penerapan prapemrosesan citra dan teks dapat membantu OCR membaca tulisan tangan serta bagaimana pengaruhnya terhadap hasil peringkasan menggunakan T5 bahasa Indonesia. Hipotesis yang digunakan adalah bahwa citra yang lebih jelas dan teks yang lebih rapi akan membantu sistem menghasilkan ringkasan yang lebih baik dibandingkan proses tanpa prapemrosesan. Oleh karena itu, tujuan penelitian ini adalah membangun dan membandingkan dua alur pengujian, yaitu alur dengan prapemrosesan dan alur tanpa prapemrosesan.

Kebaruan penelitian ini terletak pada adaptivitas *pipeline* yang diusulkan. Berbeda dari penelitian yang menerapkan rangkaian prapemrosesan secara tetap, penelitian ini tidak mengasumsikan bahwa seluruh tahap citra selalu baik untuk semua dokumen. Setiap tahapan, mulai dari *crop* area tulisan, *deskew*, *resize bicubic*, *grayscale*, *denoising*, CLAHE, *Gaussian blur*, *adaptive threshold*, hingga morfologi open, diuji sebagai kandidat dan akan dipertahankan apabila menghasilkan penurunan CER atau peningkatan akurasi karakter.

Penelitian ilmiah ini dapat diringkas dalam empat aspek. Pertama, penelitian mengusulkan *pipeline* prapemrosesan citra adaptif berbasis evaluasi OCR per dokumen. Kedua, penelitian menambahkan prapemrosesan/perbaikan teks adaptif yang hanya digunakan apabila meningkatkan skor ROUGE. Ketiga, hubungan antara kualitas OCR dan kualitas ringkasan diuji secara kuantitatif melalui korelasi akurasi/CER terhadap ROUGE. Keempat, penelitian menyajikan dokumentasi debug per tahap sehingga keputusan tahap yang dilalui atau dilewati dapat ditelusuri secara empiris.

Dengan demikian, perbedaan utama penelitian ini bukan pada penciptaan model OCR atau model transformer baru, melainkan pada mekanisme seleksi adaptif antara prapemrosesan citra, koreksi teks, dan evaluasi ringkasan. Mekanisme tersebut menjadi pembeda dari studi yang menggabungkan OCR, *image preprocessing*, dan *transformer-based summarization* tetapi masih menggunakan tahapan tetap atau hanya mengevaluasi hasil akhir tanpa menunjukkan keputusan per tahap.

2.METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen kuantitatif. Sistem dibuat untuk memproses citra tulisan tangan sampai menjadi ringkasan teks secara otomatis. Alur penelitian terdiri dari lima tahap utama, yaitu: input dokumen, prapemrosesan citra, ekstraksi teks, prapemrosesan teks, peringkasan teks, dan hasil peringkasan.

2.1. Desain, Data, dan Validasi Penelitian

Data penelitian berupa 23 file PDF tulisan tangan pada folder data yang memiliki isi dokumen sama, tetapi ditulis dengan variasi bentuk tulisan tangan dan kualitas pemindaian yang berbeda. Teknik sampling yang digunakan adalah purposive sampling, karena data dipilih sesuai kebutuhan pengujian OCR tulisan tangan dan evaluasi peringkasan. Instrumen penelitian meliputi Python, PyMuPDF untuk rendering PDF, OpenCV untuk prapemrosesan citra, PaddleOCR untuk ekstraksi teks, *LanguageTool* dan aturan koreksi sederhana untuk perbaikan teks, model T5 bahasa Indonesia untuk peringkasan, serta perhitungan CER dan ROUGE untuk evaluasi. Validasi dilakukan dengan membandingkan teks OCR terhadap transkrip referensi dan membandingkan ringkasan sistem terhadap ringkasan referensi. Setiap file diuji pada dua skenario, yaitu tanpa prapemrosesan dan dengan prapemrosesan adaptif.

Penggunaan 23 dokumen dengan isi yang sama dipilih dengan tujuan khusus, bukan sebagai data training, melainkan sebagai data testing untuk menguji dampak prapemrosesan pada model yang sudah tersedia, yaitu PaddleOCR dan T5. Dengan demikian, penelitian ini tidak menggunakan ribuan data seperti penelitian yang melatih model baru dari awal. Unit pengamatan tetap berupa dokumen utuh, dan setiap dokumen memuat teks lengkap yang dievaluasi pada level karakter. Berdasarkan transkrip referensi OCR, satu dokumen memuat sekitar 2.595 karakter setelah normalisasi, sehingga 23 dokumen setara dengan sekitar 59.685 karakter evaluasi CER. Jumlah ini memadai untuk menunjukkan tren komparatif sebelum dan sesudah prapemrosesan pada konteks studi evaluatif terbatas.

Selain itu, desain eksperimen bersifat perbandingan, karena dokumen yang sama diuji pada dua kondisi yaitu tanpa prapemrosesan dan dengan *pipeline* adaptif. Pendekatan ini mengurangi pengaruh perbedaan isi dokumen karena perbandingan dilakukan pada objek yang identik. Variasi yang diuji terutama berasal dari gaya tulisan tangan, ketebalan tinta, kemiringan, kualitas pemindaian, dan pencahayaan. Oleh karena itu, meskipun jumlah dokumen tidak besar, rancangan ini masih dapat dipertanggungjawabkan untuk menguji apakah prapemrosesan memberikan dampak terhadap OCR dan peringkasan pada kumpulan data yang ada.

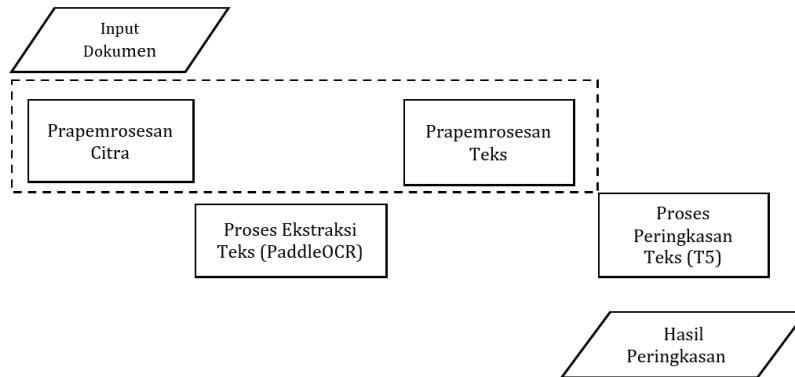
2.2. Timeline Penelitian

Penelitian dilaksanakan selama Februari sampai Juni 2026. Timeline berikut disusun berdasarkan tahapan kegiatan penelitian, mulai dari perencanaan dan studi literatur, implementasi sistem OCR dan prapemrosesan, pengujian, penyusunan manuskrip, sampai proses submit dan korespondensi awal jurnal.

Tabel 1. *Timeline Penelitian*

Bulan	Februari	Maret	April	Mei	Juni
Tahap Kegiatan	Perencanaan & Studi Literatur	Implementasi Prapemrosesan citra & OCR	Implementasi Prapemrosesan Teks & NLP	Eksperimen Pengujian & Analisis	Penyusunan Manuskrip Jurnal & Submit

Setiap bulan memiliki tahap kegiatan utama dan detail aktivitas spesifik yang menjadi dasar pelaksanaan eksperimen OCR, prapemrosesan citra, prapemrosesan teks, evaluasi CER, evaluasi ROUGE, dan penyusunan naskah jurnal.



Gambar 1. Tahapan Penelitian

Berikut adalah uraian dari masing-masing tahapan yang dilakukan dalam penelitian ini:

Input Dokumen

Tahap pertama adalah memasukkan data berupa citra digital tulisan tangan ke dalam sistem. Data diperoleh dengan mengumpulkan hasil tulisan tangan dari beberapa orang dengan konten yang sama, sehingga dalam penilaiannya akan lebih mudah dan terukur. Citra tersebut menjadi data awal yang akan diproses pada tahap pengolahan citra sebelum dibaca oleh OCR.

Prapemrosesan Citra (*Image Preprocessing*)

Citra tulisan tangan memiliki kualitas yang berbeda-beda, seperti pencahayaan tidak merata, tulisan terlalu tipis, ukuran tulisan kecil, derau latar belakang atau kemiringan halaman. Pada program, setiap tahap prapemrosesan diuji sebagai kandidat tahapan yang akan digunakan. Tahap tersebut hanya dipakai sebagai masukan tahap berikutnya apabila hasil OCR menunjukkan peningkatan akurasi, sedangkan tahap yang tidak meningkatkan akurasi dicatat dan tidak digunakan pada hasil akhir adaptif.

Tahapan prapemrosesan citra yang digunakan adalah sebagai berikut:

1. *Crop area tulisan*: Sistem memotong area yang berisi teks dan membuang bagian kosong di sekitarnya agar mesin OCR dapat lebih fokus pada konten utama.
2. *Deskew*: Kemiringan citra tulisan diperbaiki untuk menyejajarkan baris teks, sehingga karakter di dalamnya lebih mudah dibaca oleh sistem.
3. *Resize bicubic*: Citra yang telah lurus diperbesar dua kali lipat menggunakan interpolasi bicubic guna mempertegas guratan tinta tulisan tangan.
4. *Grayscale*: Tahap selanjutnya adalah mengubah citra menjadi keabuan (*grayscale*). Langkah ini berguna untuk membuang informasi warna yang tidak relevan sambil tetap mempertahankan kontras antara tinta dan latar belakang.
5. *Denoising*: Citra dibersihkan menggunakan *fastNIMeans Denoising* ($h=12$, *template WindowSize*=7, *search WindowSize*=21) untuk mereduksi bintik derau tanpa merusak bentuk huruf.
6. *CLAHE*: citra hasil *denoising* diberi *Contrast Limited Adaptive Histogram Equalization* dengan *clipLimit*=2,0 dan *tileGridSize*=(8, 8). Proses ini meningkatkan kontras lokal pada tulisan yang tipis atau redup.
7. *Gaussian blur*: citra hasil CLAHE dihaluskan menggunakan kernel 3x3. Proses ini mengurangi perubahan piksel yang terlalu tajam sebelum *thresholding*.

8. *Adaptive threshold*: citra hasil *Gaussian blur* dibinerisasi menggunakan *ADAPTIVE_THRESH_GAUSSIAN_C* dengan *blockSize=31* dan *C=15*. Proses ini memisahkan tulisan dan latar berdasarkan ambang lokal.
9. Morfologi open: citra hasil *adaptive threshold* diproses menggunakan *MORPH_OPEN* dengan kernel 2x2 satu iterasi. Proses ini membersihkan bintik kecil yang tersisa dan menghasilkan citra akhir prapemrosesan.

Ekstraksi Teks menggunakan PaddleOCR

Setelah citra diproses, ekstraksi teks dilakukan menggunakan PaddleOCR. Model ini mendeteksi area teks pada citra, mengenali karakter di dalamnya, lalu mengubahnya menjadi teks digital. Hasil dari tahap ini masih dapat mengandung kesalahan karena bentuk tulisan tangan setiap orang berbeda-beda.

Prapemrosesan Teks (Text Cleaning)

Teks hasil OCR biasanya masih mengandung salah eja, spasi berlebih, atau tanda baca yang tidak rapi. Pada tahap ini, teks dibersihkan agar lebih mudah dibaca oleh model bahasa. Prosesnya meliputi penghapusan karakter yang tidak relevan, normalisasi spasi, merapikan tanda baca dan koreksi kata secara dinamis. Pada sisi teks, perbaikan juga dibuat adaptif berdasarkan evaluasi ROUGE. Setelah OCR, teks terlebih dahulu diperbaiki dengan merapikan spasi dan tanda baca. Koreksi kata setelah OCR dilakukan secara dinamis memanfaatkan kosakata, pendekatan kemiripan kata, dan pemeriksaan ejaan. Perbaikan kalimat digunakan sebagai opsi tambahan, bukan sebagai proses yang harus dilakukan. Setiap tahapan peringkasan menggunakan model T5 dibandingkan dengan ringkasan referensi menggunakan rata-rata F1 ROUGE-1, ROUGE-2, dan ROUGE-L. Perbaikan teks hanya dipakai apabila skor gabungan ROUGE meningkat minimal 0,01 dari *baseline*. Tahapan pengoreksian teks akan dilakukan saat memang terbukti meningkatkan kualitas ringkasan, tetapi akan diabaikan jika menurunkan kualitas ringkasan.

Peringkasan Teks menggunakan Model T5

Teks yang sudah dibersihkan kemudian diringkas menggunakan model T5 (*Text-to-Text Transfer Transformer*) yang telah dilatih untuk bahasa Indonesia. Model ini membaca isi teks hasil OCR dan menghasilkan ringkasan abstraktif dalam bentuk paragraf yang lebih singkat.

Evaluasi

Evaluasi dilakukan dengan membandingkan hasil ekstraksi karakter dan hasil ringkasan dari alur yang menggunakan prapemrosesan dan alur tanpa prapemrosesan. Kualitas ekstraksi karakter diukur menggunakan CER (*Character Error Rate*), CER dihitung berdasarkan jumlah karakter yang disubstitusi (S), dihapus (D), dan disisipkan (I) dibandingkan dengan total karakter referensi asli (N)
$$CER = \frac{S + D + I}{N}$$
 sedangkan kualitas ringkasan diukur menggunakan ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), terutama ROUGE-1, ROUGE-2, dan ROUGE-L.

3.HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan alur sistem yang memproses gambar tulisan tangan menjadi teks digital, kemudian merangkumnya secara otomatis. Pembahasan disusun berdasarkan hasil pada setiap tahap sistem dan evaluasi ekstraksi karakter menggunakan CER serta evaluasi ringkasan menggunakan ROUGE.

3.1. Input Dokumen

Pada tahap awal, sistem menerima citra digital berisi tulisan tangan sebagai masukan. Citra yang diperoleh dengan mengumpulkan hasil tulisan tangan dari beberapa orang dengan konten yang sama memiliki kualitas yang beragam, baik dari pencahayaan, ketebalan tinta, maupun gaya tulisan. Kondisi ini menunjukkan mengapa prapemrosesan diperlukan sebelum OCR dilakukan.

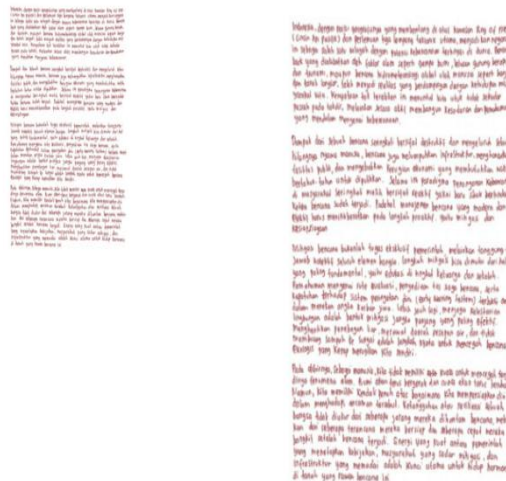
3.2. Prapemrosesan Citra (*Image Preprocessing*)

Subbab ini menampilkan hasil prapemrosesan citra sesuai tahapan yang digunakan pada program. Setiap tahap menghasilkan gambar perbandingan sebelum dan sesudah proses, sehingga perubahan visual dapat diamati secara bertahap.

Pemotongan area tulisan untuk mengurangi area kosong pada halaman.

Sebelum crop tulisan

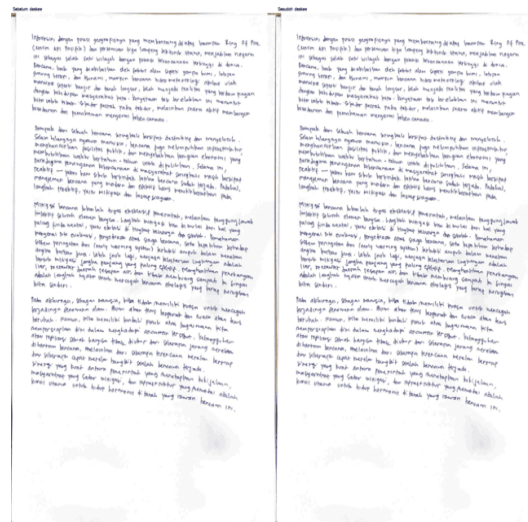
Sesudah crop tulisan



Gambar 2. Visualisasi perbandingan citra sebelum dan sesudah dipotong

Deskew

Pada tahap ini dilakukan perbaikan kemiringan citra tulisan. Proses ini membuat baris tulisan lebih sejajar sehingga karakter lebih mudah dikenali



Gambar 3. Visualisasi perbandingan citra sebelum dan sesudah dilakukan perbaikan kemiringan

Resize Bicubic

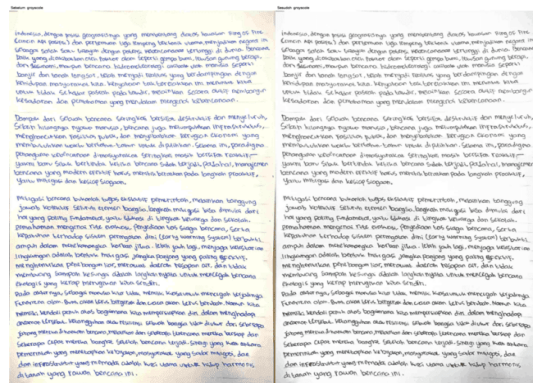
Pada tahap ini citra diperbesar dua kali lipat menggunakan interpolasi bicubic. Setelah diperbesar, goresan tulisan terlihat lebih jelas dan bentuk huruf lebih mudah diamati. Tahap ini membantu ketika ukuran tulisan pada gambar terlalu kecil.



Gambar 4. Visualisasi perbandingan citra sebelum dan sesudah resize bicubic

Konversi Grayscale

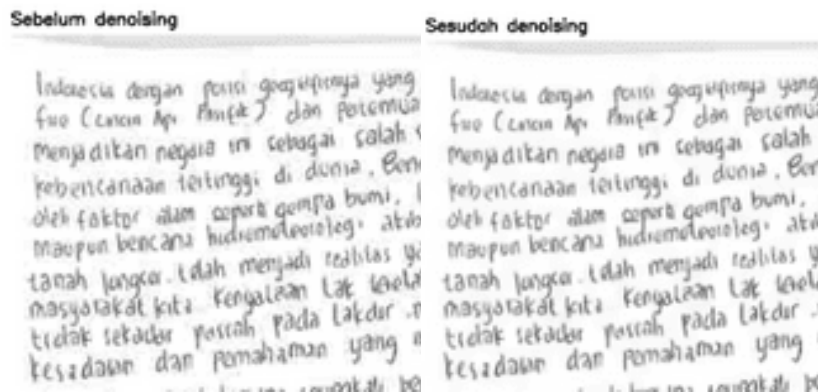
Citra hasil *resize* kemudian diubah menjadi *grayscale*. Pada tahap ini, informasi warna yang tidak diperlukan dihilangkan, sedangkan perbedaan intensitas antara tinta dan latar belakang tetap dipertahankan.



Gambar 5. Visualisasi perbandingan citra sebelum dan sesudah konversi grayscale

Denoising

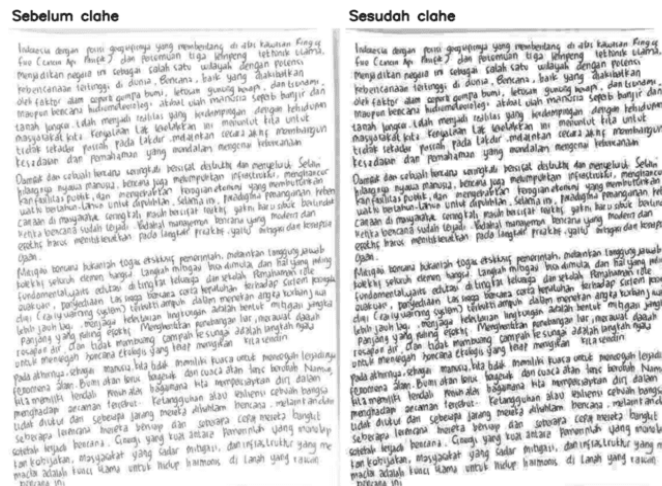
Tahap denoising mengurangi bercak kecil dan derau pada latar belakang. Hasilnya, citra terlihat lebih bersih, tetapi garis utama tulisan tetap dipertahankan agar karakter tidak berubah terlalu jauh.



Gambar 6. Visualisasi perbandingan citra sebelum dan sesudah denoising

CLAHE

CLAHE meningkatkan kontras pada area citra. Tulisan yang sebelumnya tampak tipis atau kurang jelas menjadi lebih mudah terlihat, terutama pada bagian gambar dengan pencahayaan yang tidak merata.



Gambar 7. Perbandingan citra sebelum dan sesudah CLAHE

Gaussian Blur

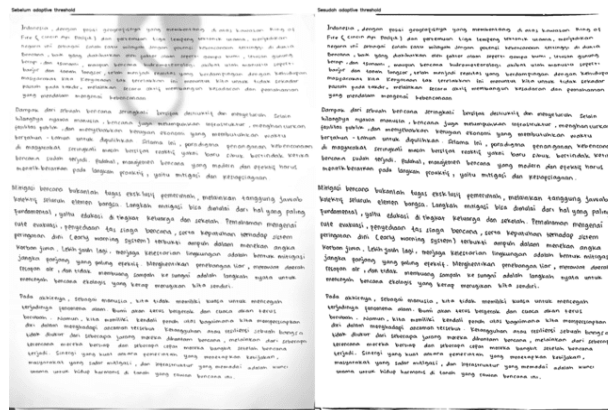
Gaussian *blur* digunakan untuk menghaluskan perubahan piksel yang terlalu tajam setelah peningkatan kontras. Tahap ini membantu mengurangi sisa derau sebelum citra masuk ke proses *adaptive threshold*.



Gambar 8. Perbandingan citra sebelum dan sesudah Gaussian blur

Adaptive Threshold

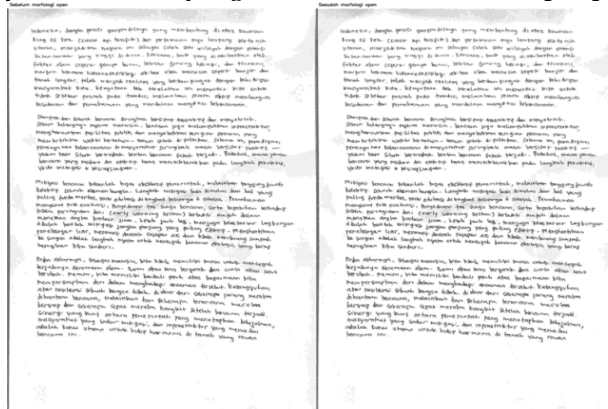
Adaptive threshold menghasilkan citra biner yang memisahkan tulisan dari latar belakang, teknik ini lebih sesuai untuk gambar tulisan tangan yang pencahayaannya tidak seragam.



Gambar 9. Visualisasi perbandingan citra sebelum dan sesudah adaptive threshold

Morfologi Open

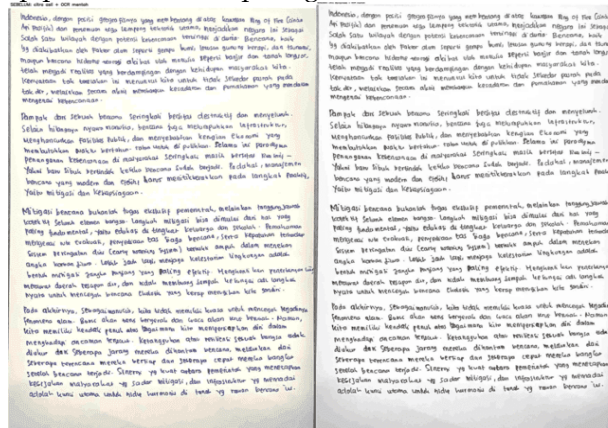
Operasi morfologi open diterapkan untuk membersihkan bintik kecil yang masih tersisa setelah *thresholding*. Hasil akhirnya adalah citra yang lebih bersih dan lebih siap diproses oleh PaddleOCR.



Gambar 10. Visualisasi perbandingan citra sebelum dan sesudah morfologi open

Hasil Perbandingan Prapemrosesan Citra Keseluruhan

Perbandingan keseluruhan menunjukkan perbedaan antara citra asli dan citra akhir yang dipilih oleh *pipeline* prapemrosesan adaptif. Secara visual, citra akhir yang terpilih umumnya memiliki kontras yang lebih jelas dan derau yang lebih rendah, tetapi sistem tetap dapat melewati tahap tertentu apabila tahap tersebut tidak meningkatkan akurasi OCR. Hasil ini mendukung fokus penelitian bahwa kualitas citra berpengaruh terhadap proses OCR dan peringkasan teks.



Gambar 11. Visualisasi perbandingan citra asli dan citra akhir hasil prapemrosesan adaptif

3.3. Hasil Ekstraksi Teks (*PaddleOCR*)

Setelah tahap citra selesai, PaddleOCR digunakan untuk mengekstraksi teks. Hasil pembacaan menunjukkan bahwa sebagian besar karakter dapat dikenali, walaupun masih terdapat kesalahan pada beberapa kata karena bentuk tulisan tangan tidak selalu konsisten.

Indonesia, dengan Posisi geografisnya, yang memberikan di atas Kanason Ring of Fire (Gincin Api pasnik) dan Perlemuan tiga lempeng tektonik ufama, menjadi kan regara in Sebagai Salah Satu Mlayah dengan polensi Kebencanaan teringgi ci dunia. Bencara back yang diakibatkan deh faktor alam seperh gempa bumi, letusan gunung berapi, dan tsunami, mau pun bencana hidrometeorologi atibal ulah manusia seperh banjir dan tanah longsor, tebah menjad realitas yang berdampingan dengan kehidupan mas yaraka lita. Kenyataan tak terebakon ini menuntut kita utuk tidak sekdar Pasrah pada takdir, melainkan Se Cara aktif mem bangun Keso daran dan femahaman yang mendalam mengenai Kebenanaan. Dampak dari Sebuah bencana seringkali bersifat destruktif dan menyeluruh Selain hilangnya nyawa manusia, bencana juga melumpuhkan infrastruktur fur, menghancurkan fasilitas publik, dan menyebabkan Kerugian ekonomi yang membutuhkan wakh. bertahun-tahununtuk dipulihkar. Selamai paracima Penanganon Kebeno di masyarakat Seringkali masih bersifat reaktif yakni baru sibuk berindak Ketika bencana Sudah terjadi. Padahal, manajemen bencana yang modern dan efektif harus menikberatkan pada langkah proaktif, yaitu mihgasi dan Kesiapsiagaan mikigasi bencana bukanlah tugas eksklusif pemerintah melainkan tanggung. Jawab Kolektiv Seluruh elemen bangsa, langkah Mikigasi bisa dimulai dari hal yang paling fundamental, yaitu edukasi di tingkat keluarga dan sekolah. Pem ahaman mengenai rute evakuasi, penyediaan tas siaga bencana, serta Kepatuhan terhadap sistem peringatan dini (early Warning System) terbukti amplit dalam menekan angka Korban Jiwa. Lebih jauh lagi, menjaga Kelestarian lingkungan adalah bentuk mikigasi Jangka panjang yang paling efektif. Menghentikan penebangan liar, merawat daerah resapan air, dan tidak Trnmbuany Sampah ke sungai adalah langkah nyata untuk mencegah bencana ekologis yang Kerap merugikan lita sendiri. Pada akhirnya, Sebagai manusia, kita tidak memiliki rese kuasa untuk mencegah ferga dinya fenomena alam. Bumi akan terus bergerak dan cuaca akan terus berubah. Namun, kita memiliki Kendali gnuh atas bagaimana Kita mempersiapkan diri dalam menghadapi ancaman tersebut. Ketangguhan atov resiliensi Sebuah bangsa tidak diukur dari seberapa Jatang mereka dihantam bencana, mebin Kan dan seberapa terencana Mereka bersipg dan seberapa cepat mereka bangkit setelah bencana terjadi. Sinergi yang kuat antara pemerintah yang menetapkan kebijakan, masyarakat yang sadar mihgasi, dan infrastruktur yang memadai adalah Kunci utama untuk hidup harmonis di tanah yang rawan bencana ini.

Gambar 12. Visualisasi hasil ekstraksi teks oleh PaddleOCR

Hasil Prapemrosesan Teks (*Text Cleaning*)

Teks hasil OCR biasanya masih mengandung salah eja, spasi berlebih, atau tanda baca yang tidak rapi. Pada tahap ini, teks dibersihkan agar lebih mudah dibaca oleh model bahasa. Prosesnya meliputi penghapusan karakter yang tidak relevan, normalisasi spasi, merapikan tanda baca, dan mengoreksi kata serta kalimat.

(kanan api pasnik) dan pertemuan tiga lempeng tektonik utama, menjadikan negara in sebagai salah satu wilayah dengan potensi Kebencanaan tertinggi; di dunia. Bencana, baik yang diakibatkan oleh faktor alam seperti gempa bumi letusan gunung berapi, dan tsunami, maupun bencana hidrometeorologi akibat ulah manusia seperti banjir dan tanah longsor, telah menjadi realitas yang berdampingan dengan kehidupan masyarakat kita, kenyataan tak terelakkan ini menuntut kita untuk tidak sekedar pasrah pada takdir, melainkan secara aktif membangun kesadaran dan pemaha- man yang mendalam mengenai Perencanaan. Dampak dari sebuah bencana Seringkali bersifat destruktif dan menyeluruh Selain hilangnya nyawa manusia, bencana juga melumpuhkan infrastruktur, menghancurkan Fasilitas publik, dan menyebabkan kerugian ekonomi yang membutuhkan waktu bertahun-tahun untuk dipulihkan. Selama ini, paradigma penanganan kebencanaan di masyarakat seringkali masih bersifat reaktif. Yaitu baru sibuk bertindak ketika bencana sudah terjadi. Padahal, manajemen bencana yang modern dan efektif harus menitikberatkan pada langkah proaktif, yaitu mitigasi dan kesiapsiagaan. Mitigasi bencana beaunan tugas eksklusif pemerintah, melainkan tanggung Jawab Kolektiv seluruh elemen bangsa. Langkah mitigasi bisa dimulai dari hal yang paling Fundamental, yaitu edukasi di tingkat keluarga dan sekolah Pemahaman mengenai nee evakuasi, penyediaan tas siaga rencana, serta kepatuhan terhadap sistem peringatan dini (early warning system) terbukti ampu dalam menekan angka korban jiwa. Lebih jauh lagi, menjaga kelestarian lingkungan adalah bentuk mitigasi jangka panjang yang paling efektif. Menghentikan penebangan liar, merawat daerah resapan air, dan tidak membuang sampah ke sungai adalah langkah nyata untuk mencegah bencana ekologis yang kerap merugikan kita sendiri. Pada akhirnya, sebagai manusia, kita tidak memiliki kuasa untuk mencegah terjal- dinya fenomena alam. Bumi akan terus bergerak dan cuaca akan terus berubah. Namun, Kita memiliki kendali penuh atas bagaimana kita mempersiapkan diri dalam menghadapi ancaman tersebut. Ketangguhan atau resiliensi sebuah bangsa tidak diukur dari seberapa Jatang mereka dihantam bencana, melainkan dari seberapa terencana mereka bersipg dan seberapa cepat mereka bangkit setelah bencana terjadi Sinergi yang kuat antara pemerintah yang menetapkan kebijakan, masyarakat yang sadar mitigasi, dan inprastuuk- tur yang memadai adalah kunci utama untuk hidup harmonis di tanah yang rawan bencana ini.

(kanan api pasnik) dan pertemuan tiga lempeng tektonik utama, menjadikan negara in sebagai salah satu wilayah dengan potensi Kebencanaan tertinggi; di dunia. Bencana, baik yang diakibatkan oleh faktor alam seperti gempa bumi letusan gunung berapi, dan tsunami, maupun bencana meteorologi akibat ulah manusia seperti banjir dan tanah longsor, telah menjadi realitas yang berdampingan dengan kehidupan masyarakat kita, kenyataan tak terelakkan ini menuntut kita untuk tidak sekedar pasrah pada takdir, melainkan secara aktif membangun kesadaran dan pemahat - man yang mendalam mengenai Perencanaan. Dampak dari sebuah bencana Seringkali bersifat destruktif dan menyeluruh Selain hilangnya nyawa manusia, bencana juga melumpuhkan infrastruktur, menghancurkan Fasilitas publik, dan menyebabkan kerugian ekonomi yang membutuhkan waktu bertahun-tahun untuk dipulihkan. Selama ini, paradigma penanganan perencanaan di masyarakat seringkali masih bersifat reaktif. Yaitu baru sibuk bertindak ketika bencana sudah terjadi. Padahal, manajemen bencana yang modern dan efektif harus menitikberatkan pada langkah proaktif, yaitu mitigasi dan kesiapsiagaan. Mitigasi bencana beaunan tugas eksklusif pemerintahan, melainkan tanggung Jawab Kolektiv seluruh elemen bangsa. Langkah mitigasi bisa dimulai dari hal yang paling Fundamental, yaitu edukasi di tingkat keluarga dan sekolah Pemahaman mengenai nee evakuasi, penyediaan tas siaga rencana, serta kepatuhan terhadap sistem peringatan dini (early warning system) terbukti ampu dalam menekan angka korban jiwa. Lebih jauh lagi, menjaga kelestarian lingkungan adalah bentuk mitigasi jangka panjang yang paling efektif. Menghentikan penebangan liar, merawat daerah resapan air, dan tidak membuang sampah ke sungai adalah langkah nyata untuk mencegah bencana ekologis yang kerap merugikan kita sendiri. Pada akhirnya, sebagai manusia, kita tidak memiliki kuasa untuk mencegah terjal - dinya fenomena alam. Bumi akan terus bergerak dan cuaca akan terus berubah. Namun, Kita memiliki kendali penuh atas bagaimana kita mempersiapkan diri dalam menghadapi ancaman tersebut. Ketangguhan atau resensi sebuah bangsa tidak diukur dari seberapa Jatang mereka dihantam bencana, melainkan dari seberapa terencana mereka bersipg dan seberapa cepat mereka bangkit setelah bencana terjadi Sinergi yang kuat antara pemerintah yang menetapkan kebijakan, masyarakat yang sadar mitigasi, dan inprastuuk - tur yang memadai adalah kunci utama untuk hidup harmonis di tanah yang rawan bencana ini.

Gambar 13. Perbandingan teks sebelum dan sesudah pembersihan hasil OCR

Hasil Peringkasan Teks (*Model T5*)

Tahap akhir adalah peringkasan teks menggunakan model T5. Model membaca teks yang telah dibersihkan, lalu menghasilkan ringkasan yang lebih singkat. Ringkasan yang dihasilkan tidak hanya mengambil potongan kalimat, tetapi menyusun kembali informasi utama dalam bentuk paragraf baru.

Bencana yang diakibatkan oleh faktor alam seperti gempa bumi, letusan gunung berapi, dan tsunami, maupun bencana meteorologi akibat ulah manusia seperti banjir dan tanah longsor, telah menjadi kenyataan yang berdampak dengan kehidupan masyarakat kita. Kenyataan tersebut menuntut kita untuk tidak sekadar pasrah pada takdir, melainkan mencari cara yang tepat untuk membangun kesadaran dan memohon yang mendalam mengenai kebenaran. Ketangguhan suatu bangsa tidak diukur dari seberapa sering mereka dihantam bencana, melainkan seberapa terencana mereka bersiap dan seberapa cepat mereka bangkit setelah bencana terjadi. Sinergi yang kuat antara pemerintah yang menetapkan kebijakan, masyarakat yang sadar mitigasi, dan infrastruktur yang memadai adalah kunci utama untuk hidup harmonis di tanah yang rawan bencana ini.

Gambar 14. Contoh hasil ringkasan

Evaluasi Ekstraksi Karakter (CER) dan Peringkasan Teks (ROUGE Score)

Perbandingan dilakukan antara alur tanpa prapemrosesan citra dan alur dengan prapemrosesan citra adaptif. Pada alur adaptif, setiap tahap citra diuji terhadap hasil OCR sementara, tahapan hanya dipakai bila CER menurun atau akurasi karakter meningkat. Untuk evaluasi ringkasan, perbandingan juga memperhatikan prapemrosesan/perbaikan teks sebelum model T5 menghasilkan ringkasan.

Evaluasi Character Error Rate (CER)

digunakan untuk mengukur tingkat kesalahan karakter hasil OCR terhadap teks referensi. Rumus yang digunakan adalah $CER = (S + D + I) / N$. (S) adalah jumlah substitusi karakter, (D) adalah jumlah penghapusan karakter, (I) adalah jumlah penyisipan karakter dan (N) adalah jumlah karakter pada teks referensi. Nilai akurasi karakter dihitung dengan rumus $Akurasi = (1 - CER) \times 100\%$.

Contoh pada data 1.pdf, sebelum prapemrosesan citra diperoleh CER 0.3102. Jika dinyatakan per 1.000 karakter referensi, jumlah kesalahan setara dengan $0.3102 \times 1.000 = 310.21$ karakter, sehingga akurasi = $(1 - 0.3102) \times 100\% = 68.98\%$. Setelah prapemrosesan citra, CER menjadi 0.2131, jumlah kesalahan setara dengan 213.10 karakter per 1.000 karakter referensi, dan akurasi menjadi 78.69%.

Tabel 2. Hasil evaluasi CER dan akurasi OCR sebelum dan setelah prapemrosesan citra

File	Sebelum Prapemrosesan		Setelah Prapemrosesan	
	CER	Akurasi (%)	CER	Akurasi (%)
1.pdf	0.3102	68.98	0.2131	78.69
2.pdf	0.0597	94.03	0.0563	94.37
3.pdf	0.2112	78.88	0.1495	85.05
4.pdf	0.0809	91.91	0.0809	91.91
5.pdf	0.0898	91.02	0.0898	91.02
6.pdf	0.3530	64.70	0.1175	88.25
7.pdf	0.2085	79.15	0.2042	79.58
8.pdf	0.5056	49.44	0.4123	58.77
9.pdf	0.4501	54.99	0.3653	63.47
10.pdf	0.3530	64.70	0.2979	70.21
11.pdf	0.2821	71.79	0.2208	77.92
12.pdf	0.2405	75.95	0.2389	76.11
13.pdf	0.7599	24.01	0.4374	56.26
14.pdf	0.6971	30.29	0.5715	42.85
15.pdf	0.4482	55.18	0.2470	75.30
16.pdf	0.1179	88.21	0.1129	88.71
17.pdf	0.8224	17.76	0.5341	46.59
18.pdf	0.2301	76.99	0.2108	78.92
19.pdf	0.2216	77.84	0.1468	85.32
20.pdf	0.3796	62.04	0.2886	71.14
21.pdf	0.1561	84.39	0.1561	84.39
22.pdf	0.3626	63.74	0.2933	70.67
23.pdf	0.6381	36.18	0.2813	71.87
Rata-rata	0.3469	65.31	0.2490	75.10

Berdasarkan rata-rata seluruh data, CER menurun dari 0.3469 menjadi 0.2490. Karena CER yang lebih kecil menunjukkan kesalahan karakter yang lebih rendah, hasil ini menunjukkan peningkatan kualitas OCR. Rata-rata akurasi karakter juga meningkat dari 65.31% menjadi 75.10%, atau naik sebesar 9.79 poin persentase.

Analisis per tahap menunjukkan bahwa peningkatan akurasi tidak berasal dari satu tahap yang selalu dominan pada semua dokumen. Secara rata-rata, morfologi open menghasilkan akurasi tahap sebesar 71.80% dan *Gaussian blur* 71.57%, sedangkan *CLAHE* menghasilkan 68.34% atau naik 3.02 poin dari *baseline*. *Adaptive threshold* tidak selalu meningkatkan rata-rata apabila dipaksakan ke seluruh dokumen, tetapi dapat memberi pengaruh besar pada dokumen tertentu, misalnya data 6.pdf yang melewati tahap *grayscale*, *denoising*, *CLAHE*, dan *adaptive threshold* sehingga akurasi naik dari 64.70% menjadi 88.25%. Temuan ini memperkuat alasan penggunaan *pipeline* adaptif bahwa tahap yang baik pada satu dokumen belum tentu baik pada dokumen lain.

Secara teknis, *CLAHE* membantu ketika tulisan memiliki kontras lokal yang lemah, tinta terlihat tipis, atau pencahayaan tidak merata. Proses ini memperkuat perbedaan intensitas antara guratan tinta dan latar belakang sehingga deteksi karakter oleh OCR menjadi lebih stabil. *Adaptive threshold* berperan ketika latar halaman tidak seragam karena ambang biner dihitung secara lokal, bukan global. Akan tetapi, pada dokumen yang sejak awal sudah bersih atau memiliki guratan tebal, *thresholding* dapat menghilangkan detail halus karakter. Karena itu, *adaptive threshold* sebaiknya tidak dipakai secara otomatis untuk semua dokumen, tetapi dipilih berdasarkan dampaknya terhadap metrik OCR.

Perbedaan peningkatan antar dokumen juga dipengaruhi oleh kualitas awal dokumen. Peningkatan tertinggi terjadi pada dokumen dengan akurasi awal rendah, misalnya 23.pdf naik 35.68 poin, 13.pdf naik 32.25 poin, 17.pdf naik 28.82 poin, dan 6.pdf naik 23.55 poin. Dokumen ini lebih banyak memperoleh peningkatan poin dari *resize bicubic*, *denoising*, *CLAHE*, *Gaussian blur*, *adaptive threshold*, atau morfologi open karena *baseline* OCR masih terganggu oleh ukuran tulisan, kontras, derau, atau pemisahan latar yang kurang baik. Sebaliknya, 4.pdf, 5.pdf, dan 21.pdf hampir tidak berubah karena akurasi awalnya sudah tinggi, *pipeline* adaptif kemudian cenderung mempertahankan hasil asli dan melewati tahap yang tidak memberi peningkatan.

Untuk memperlihatkan sifat adaptif *pipeline* secara lebih eksplisit, Tabel 3 menyajikan tahapan prapemrosesan yang benar-benar dilalui oleh setiap file. Kolom tahapan dilalui menunjukkan tahap yang diterima karena memberikan perbaikan terhadap hasil OCR, sedangkan kolom tahapan dilewati menunjukkan tahap yang tidak dipertahankan karena tidak meningkatkan metrik pada file tersebut.

Tabel 3. Tahapan prapemrosesan citra adaptif yang dilalui setiap file

File	Tahapan dilalui/dipakai	Tahapan dilewati	Tahap OCR terbaik	Delta akurasi (poin)
1.pdf	resize_bicubic	crop_tulisan > deskew > grayscale > denoising > clahe > gaussian_blur > adaptive_threshold >...	resize_bicubic	9.71
2.pdf	grayscale > clahe	crop_tulisan > deskew > resize_bicubic > denoising > gaussian_blur > adaptive_threshold >...	clahe	0.35
3.pdf	crop_tulisan > resize_bicubic	deskew > grayscale > denoising > clahe > gaussian_blur > adaptive_threshold > morfologi_open	resize_bicubic	6.17
4.pdf	-	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > clahe > gaussian_blur > ...	asli	0.00
5.pdf	-	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > clahe > gaussian_blur > ...	asli	0.00
6.pdf	grayscale > denoising > clahe >	crop_tulisan > deskew > resize_bicubic > gaussian_blur > morfologi_open	adaptive_thresh old	23.55

	adaptive_thresho ld			
7.pdf	grayscale	crop_tulisan > deskew > resize_bicubic > denoising > clahe > gaussian_blur > adaptive_threshold...	grayscale	0.42
8.pdf	grayscale	crop_tulisan > deskew > resize_bicubic > denoising > clahe > gaussian_blur > adaptive_threshold...	grayscale	9.33
9.pdf	resize_bicubic > clahe > gaussian_blur > morfologi_open	crop_tulisan > deskew > grayscale > denoising > adaptive_threshold	morfologi_open	8.48
10.pdf	denoising > morfologi_open	crop_tulisan > deskew > resize_bicubic > grayscale > clahe > gaussian_blur > adaptive_threshold...	morfologi_open	5.51
11.pdf	resize_bicubic > grayscale > morfologi_open	crop_tulisan > deskew > denoising > clahe > gaussian_blur > adaptive_threshold	morfologi_open	6.13
12.pdf	grayscale	crop_tulisan > deskew > resize_bicubic > denoising > clahe > gaussian_blur > adaptive_threshold...	grayscale	0.15
13.pdf	resize_bicubic > grayscale > denoising > clahe > morfologi_open	crop_tulisan > deskew > gaussian_blur > adaptive_threshold	morfologi_open	32.25
14.pdf	grayscale > denoising > morfologi_open	crop_tulisan > deskew > resize_bicubic > clahe > gaussian_blur > adaptive_threshold	morfologi_open	12.56
15.pdf	grayscale > denoising	crop_tulisan > deskew > resize_bicubic > clahe > gaussian_blur > adaptive_threshold > mor...	denoising	20.12
16.pdf	gaussian_blur	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > clahe > adaptive_thresholdesho...	gaussian_blur	0.50
17.pdf	resize_bicubic > denoising > clahe > gaussian_blur	crop_tulisan > deskew > grayscale > adaptive_threshold > morfologi_open	gaussian_blur	28.82
18.pdf	clahe > morfologi_open	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > gaussian_blur > adaptive_threshold.	morfologi_open	1.93
19.pdf	denoising > gaussian_blur	crop_tulisan > deskew > resize_bicubic > grayscale > clahe > adaptive_threshold > morfologi_open	gaussian_blur	7.48
20.pdf	gaussian_blur > morfologi_open	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > clahe > adaptive_threshold	morfologi_open	9.09
21.pdf	-	crop_tulisan > deskew > resize_bicubic > grayscale > denoising > clahe > gaussian_blur > ...	asli	0.00
22.pdf	grayscale > denoising	crop_tulisan > deskew > resize_bicubic > clahe > gaussian_blur > adaptive_threshold > morfologi_open	denoising	6.94
23.pdf	crop_tulisan > resize_bicubic > gaussian_blur	deskew > grayscale > denoising > clahe > adaptive_threshold > morfologi_open	gaussian_blur	35.68

Tabel 3 menunjukkan bahwa pipeline tidak menerapkan tahapan yang sama pada semua file. Sebagai contoh, 4.pdf, 5.pdf, dan 21.pdf mempertahankan citra asli karena prapemrosesan tidak

meningkatkan akurasi, sedangkan 6.pdf menggunakan kombinasi *grayscale*, *denoising*, *CLAHE*, dan *adaptive threshold* dengan kenaikan akurasi 23.55 poin. File 23.pdf menggunakan *crop* tulisan, *resize bicubic*, dan *Gaussian blur* dengan kenaikan 35.68 poin. Variasi ini memperlihatkan bahwa keputusan prapemrosesan bersifat adaptif terhadap karakteristik masing-masing dokumen, bukan berupa pipeline tetap yang dipaksakan pada seluruh data.

Evaluasi ROUGE

ROUGE digunakan untuk mengukur kemiripan ringkasan sistem terhadap ringkasan referensi. ROUGE-1 menghitung kesesuaian unigram, ROUGE-2 menghitung kesesuaian bigram, sedangkan ROUGE-L menggunakan *longest common subsequence*. Nilai yang ditampilkan pada tabel adalah F1-score.

$Precision = \text{jumlah } n\text{-gram yang tumpang tindih} / \text{jumlah } n\text{-gram pada ringkasan sistem}$

$Recall = \text{jumlah } n\text{-gram yang tumpang tindih} / \text{jumlah } n\text{-gram pada ringkasan referensi}$

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

Contoh pada data 1.pdf, ROUGE-1 sebelum prapemrosesan/perbaikan teks adalah 0.0876 atau 8.76%, sedangkan setelah prapemrosesan/perbaikan teks menjadi 0.2955 atau 29.55%. ROUGE-2 meningkat dari 0.0321 menjadi 0.1298, dan ROUGE-L meningkat dari 0.0797 menjadi 0.2727. Mode perbaikan teks adaptif yang dipilih pada data ini adalah koreksi_kata.

Tabel 4. Hasil evaluasi ROUGE sebelum dan setelah prapemrosesan/perbaikan teks

File	Sebelum perbaikan teks dan kalimat			Setelah Perbaikan teks atau kalimat		
	R1	R2	RL	R1	R2	RL
1.pdf	0.0876	0.0321	0.0797	0.2955	0.1298	0.2727
2.pdf	0.3651	0.1680	0.2460	0.4566	0.2949	0.3470
3.pdf	0.2033	0.0328	0.1545	0.4179	0.1729	0.3433
4.pdf	0.3782	0.1525	0.2857	0.4819	0.2591	0.3775
5.pdf	0.4113	0.2114	0.3548	0.4286	0.2720	0.3889
6.pdf	0.4215	0.2583	0.3802	0.5381	0.3891	0.4753
7.pdf	0.2066	0.0750	0.1405	0.3596	0.1887	0.2846
8.pdf	0.0372	0.0000	0.0372	0.1057	0.0304	0.0830
9.pdf	0.0914	0.0000	0.0711	0.2845	0.0696	0.1466
10.pdf	0.1650	0.0490	0.1262	0.2550	0.0964	0.2231
11.pdf	0.3710	0.1789	0.2984	0.4062	0.2283	0.3359
12.pdf	0.1287	0.0000	0.0585	0.3843	0.1850	0.2620
13.pdf	0.0213	0.0000	0.0106	0.1370	0.0184	0.0822
14.pdf	0.0442	0.0112	0.0442	0.1273	0.0459	0.1091
15.pdf	0.3185	0.0970	0.2444	0.3000	0.0388	0.2077
16.pdf	0.3789	0.2133	0.2996	0.4255	0.3226	0.4043
17.pdf	0.0606	0.0000	0.0485	0.3053	0.1489	0.2737
18.pdf	0.1840	0.0484	0.1520	0.2281	0.0690	0.1901
19.pdf	0.4549	0.3117	0.3519	0.4206	0.2944	0.3863
20.pdf	0.1591	0.0115	0.1023	0.2199	0.0635	0.1152
21.pdf	0.3475	0.1479	0.2625	0.3959	0.2256	0.3858
22.pdf	0.0928	0.0000	0.0722	0.3500	0.1513	0.3000
23.pdf	0.1061	0.0000	0.0735	0.0920	0.0232	0.0843
Rata-rata	0.2189	0.0869	0.1693	0.3224	0.1616	0.2643

Rata-rata ROUGE-1 meningkat dari 0.2189 menjadi 0.3224, ROUGE-2 meningkat dari 0.0869 menjadi 0.1616, dan ROUGE-L meningkat dari 0.1693 menjadi 0.2643. Kenaikan masing-masing adalah 10.35, 7.47, dan 9.50 poin persentase.

Untuk menguji hubungan antara kualitas OCR dan kualitas ringkasan secara kuantitatif, dilakukan analisis korelasi pada 23 file uji. Variabel OCR diwakili oleh akurasi karakter dan CER, sedangkan kualitas ringkasan diwakili oleh ROUGE-1, ROUGE-2, dan ROUGE-L. Hasil korelasi menunjukkan bahwa pada skenario dengan prapemrosesan, akurasi OCR memiliki korelasi positif terhadap ROUGE-

1 ($r = 0.762$), ROUGE-2 ($r = 0.723$), dan ROUGE-L ($r = 0.741$), dengan p -value < 0.001 . Sebaliknya, CER berkorelasi negatif terhadap ROUGE-L ($r = -0.741$), yang berarti semakin kecil kesalahan karakter maka kecenderungan kualitas ringkasan semakin baik.

Tabel 5. Korelasi akurasi/CER OCR terhadap skor ROUGE

Skenario	Variabel	Pearson r	p	Spearman rho	p	Interpretasi
Dengan prapemrosesan	Akurasi OCR vs ROUGE-1	0.762	<0.001	0.837	<0.001	positif kuat
Dengan prapemrosesan	Akurasi OCR vs ROUGE-2	0.723	<0.001	0.807	<0.001	positif kuat
Dengan prapemrosesan	Akurasi OCR vs ROUGE-L	0.741	<0.001	0.813	<0.001	positif kuat
Dengan prapemrosesan	CER vs ROUGE-L	-0.741	<0.001	-0.813	<0.001	negatif kuat
Tanpa prapemrosesan	Akurasi OCR vs ROUGE-L	0.665	<0.001	0.736	<0.001	positif sedang-kuat
Perubahan nilai	Delta akurasi vs delta ROUGE-L	-0.160	0.466	-0.236	0.277	tidak signifikan

Hasil ini mendukung kesimpulan bahwa kualitas OCR yang lebih tinggi cenderung berasosiasi dengan kualitas ringkasan yang lebih baik. Namun, korelasi perubahan akurasi terhadap perubahan ROUGE-L tidak signifikan. Artinya, kenaikan OCR pada setiap file tidak selalu berbanding lurus secara linear dengan kenaikan ROUGE, karena ringkasan juga dipengaruhi oleh panjang teks OCR, kesalahan pada kata kunci, keputusan perbaikan teks adaptif, dan perilaku generatif model T5. Dengan demikian, peningkatan OCR merupakan faktor penting, tetapi bukan satu-satunya faktor yang menentukan kualitas ringkasan.

Untuk memperkuat argumen bahwa 23 dokumen masih dapat digunakan pada desain evaluasi ini, dilakukan pula uji komparatif perbandingan antara nilai sebelum dan sesudah prapemrosesan. Hasilnya menunjukkan bahwa peningkatan akurasi OCR, penurunan CER, serta peningkatan ROUGE secara statistik pada data uji yang digunakan.

Tabel 6. Uji komparatif perbandingan sebelum dan sesudah prapemrosesan

Metrik	Sebelum	Sesudah	Perubahan	t	p-value	Cohen dz
Akurasi OCR (%)	65.31	75.10	+9.79 poin	4.313	0.000281	0.899
CER	0.3469	0.2490	-0.0979	4.313	0.000281	0.899
ROUGE-1	0.2189	0.3224	+0.1035	5.580	0.000013	1.163
ROUGE-2	0.0869	0.1616	+0.0747	5.998	0.0000049	1.251
ROUGE-L	0.1693	0.2643	+0.0950	6.231	0.0000029	1.299

Hasil uji perbandingan ini tidak dimaksudkan untuk mengklaim generalisasi luas seperti pada penelitian *machine learning* skala besar, tetapi untuk menunjukkan validitas internal bahwa pada 23 dokumen yang sama, kondisi setelah prapemrosesan memberikan peningkatan yang konsisten dan terukur dibandingkan kondisi sebelum prapemrosesan.

Keterbatasan Penelitian dan Generalisasi

Keterbatasan utama penelitian ini tetap terletak pada jumlah dokumen yang masih terbatas, yaitu 23 dokumen tulisan tangan dengan isi yang sama. Namun, batasan tersebut perlu dipahami sesuai tujuan penelitian. Data tidak digunakan untuk melatih model baru, melainkan untuk evaluasi komparatif terhadap model yang sudah ada. Kesamaan isi dokumen justru membantu mengendalikan variabel isi teks, sehingga perbedaan hasil lebih banyak dipengaruhi oleh kualitas visual tulisan, gaya penulis, ketebalan tinta, kemiringan, noise, dan kualitas pemindaian.

Dengan demikian, 23 dokumen masih dapat dipertanggungjawabkan untuk studi awal yang berfokus pada pengujian dampak prapemrosesan secara perbandingan. Akan tetapi, generalisasi eksternal tetap perlu dilakukan secara hati-hati. Tulisan yang rapi, kontras tinggi, dan tidak miring

cenderung sudah dapat dibaca dengan baik oleh OCR sehingga peningkatan prapemrosesan menjadi kecil. Sebaliknya, tulisan yang kecil, tipis, miring, buram, atau memiliki latar tidak merata lebih berpotensi memperoleh manfaat besar dari pipeline adaptif. Penelitian lanjutan perlu memperluas data dengan variasi isi dokumen, lebih banyak penulis, lebih banyak kondisi pemindaian/foto, dan tata letak dokumen yang lebih beragam.

4.KESIMPULAN

Penelitian ini menunjukkan bahwa prapemrosesan citra adaptif memberikan dampak positif terhadap pengenalan karakter pada dokumen tulisan tangan. Berdasarkan 23 file uji, rata-rata CER menurun dari 0.3469 menjadi 0.2490, sedangkan rata-rata akurasi meningkat dari 65.31% menjadi 75.10%. Kenaikan akurasi sebesar 9.79 poin persentase menunjukkan bahwa seleksi tahap citra berbasis metrik lebih sesuai daripada memaksakan seluruh tahapan pada semua file. Dari sisi peringkasan, prapemrosesan/perbaikan teks adaptif juga memberikan peningkatan kualitas hasil ringkasan. Rata-rata ROUGE-1 meningkat sebesar 10.35 poin persentase, ROUGE-2 meningkat sebesar 7.47 poin persentase, dan ROUGE-L meningkat sebesar 9.50 poin persentase. Hasil ini memperlihatkan bahwa kualitas OCR dan kualitas teks masukan berpengaruh langsung terhadap ringkasan yang dihasilkan model T5. Rekomendasi dari penelitian ini adalah mempertahankan evaluasi prapemrosesan citra adaptif sebelum OCR, bukan memaksakan satu filter yang sama untuk semua halaman. Perbaikan teks juga disarankan tetap adaptif berdasarkan kenaikan ROUGE, serta data uji perlu diperluas dengan variasi bentuk tulisan tangan, kualitas pemindaian, dan tingkat kemiringan dokumen. Secara data, hasil ini menunjukkan bahwa pendekatan adaptif lebih tepat untuk dokumen tulisan tangan dibandingkan prapemrosesan yang harus melewati seluruh tahapan prapemrosesan. Setiap halaman dapat memiliki karakteristik tulisan tangan yang berbeda, sehingga sistem perlu mengevaluasi tahapan prapemrosesan citra dan tahapan prapemrosesan teks berdasarkan metrik yang terukur. Pemilihan tahapan prapemrosesan berbasis CER dan pemilihan perbaikan teks berbasis ROUGE membantu menjaga agar prapemrosesan hanya digunakan ketika benar-benar memberi peningkatan. Analisis tambahan menunjukkan adanya korelasi positif kuat antara akurasi OCR dan skor ROUGE pada hasil dengan prapemrosesan, khususnya ROUGE-L ($r = 0.741$; $p < 0.001$), sehingga hubungan antara peningkatan kualitas OCR dan kualitas ringkasan memperoleh dukungan empiris. Meskipun demikian, generalisasi hasil masih terbatas karena data uji hanya mencakup 23 dokumen dengan isi yang sama, penelitian lanjutan perlu melibatkan variasi penulis, isi dokumen, dan kondisi pemindaian yang lebih luas. Dengan mempertimbangkan bahwa penelitian ini merupakan evaluasi/*testing*, bukan pelatihan model baru, penggunaan 23 dokumen masih relevan untuk menunjukkan kecenderungan peningkatan secara perbandingan. Akan tetapi, klaim tetap dibatasi pada konteks dokumen uji yang digunakan dan perlu diperluas pada penelitian selanjutnya.

REFERENCES

- [1] Agustian, I. A. (2025). *Implementation of the T5-small model for abstractive text summarization in SEO meta description generation*. Universitas Nusa Putra.
- [2] Almegbil, L., & al., et. (2026). Raqim: A hybrid framework for correcting Arabic OCR errors by integrating dictionary-based and generative techniques. *Procedia Computer Science*, 275, 493–502.
- [3] Arifin, M. N., Mansyur, M. U., & dkk. (2025). Enhanced OCR recognition for Madurese text documents: A genetic algorithm approach with Tesseract 5.5. *JUITA*, 13(2).
- [4] Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). What is wrong with scene text recognition model comparisons? Dataset and model analysis. *Proc. IEEE/CVF Int. Conf. Computer Vision*, 4715–4723.
- [5] Bhaskoro, S. B., Pratama, R. A., & Darmawan, S. A. H. (2025). Deteksi dan interpretasi tulisan tangan bahasa Indonesia melalui pemrosesan citra dan optical character recognition (OCR). *JTRM (Jurnal Teknologi Dan Rekayasa Manufaktur)*, 7(1), 48–64.
- [6] Biswas, S., & Das, A. K. (2020). Effects of morphological operations on optical character recognition for degraded text images. *Journal of Visual Communication and Image Representation*, 68, 102769.
- [7] Cahyawijaya, S., & al., et. (2021). IndoNLP: Benchmark and resources for evaluating Indonesian natural language generation. *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, 8875–8898.
- [8] Dey, R., Singh, D., Balabantaray, R. C., Karuppiiah, M., Mohanty, S., & Samanta, D. (2022). Approach for preprocessing in offline optical character recognition (OCR). *Proc. Interdisciplinary Research in Technology and*

- Management (IRTM)*, 1–6.
- [9] Ingle, R. R., Fujii, Y., Deselaers, T., Baccash, J., & Popat, A. C. (2019). A scalable handwritten text recognition system. *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 17–24.
- [10] Jiao, L., & Zhang, Y. (2019). Attention-based end-to-end neural network for handwritten character recognition. *IEEE Access*, 7, 9123–9130.
- [11] Koto, F., Lau, J. H., & Baldwin, T. (2020). Liputan6: A large-scale Indonesian dataset for text summarization. *Proc. AACL-IJCNLP*, 5982–5998.
- [12] Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access*, 8, 142642–142668.
- [13] Mursari, L. R., & Wibowo, A. (2021). The effectiveness of image preprocessing on digital handwritten scripts recognition with the implementation of OCR Tesseract. *Computer Engineering and Applications Journal*, 10(3), 177–186.
- [14] Nugroho, A. S., & Hidayatullah, A. F. (2019). Analisis kinerja OCR Tesseract pada pengenalan karakter dokumen berbahasa Indonesia. *Jurnal Sistem Informasi*, 15(1), 12–20.
- [15] Purnama, I. N., & Utami, N. N. W. (2023). Implementasi peringkasan dokumen berbahasa Indonesia menggunakan metode text to text transfer transformer (T5). *Jurnal Teknologi Informasi Dan Komputer*, 9(4), 374–383.
- [16] Raffel, C., & al., et. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- [17] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, 1073–1083.
- [18] Setiawan, A., & Purnamasari, D. (2021). Text summarization pada artikel berita berbahasa Indonesia menggunakan metode transformer. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(3), 450–457.
- [19] Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2298–2304.
- [20] Singh, A., & Gupta, A. (2023). A review on preprocessing techniques for handwritten text recognition. *Multimedia Tools and Applications*, 82(10), 15555–15578.
- [21] Syafiq, M. A., & Adiwijaya, A. (2023). Kinerja model bahasa berbasis transformer untuk peringkasan teks abstraktif pada dokumen bahasa Indonesia. *Jurnal Media Informatika Budidarma*, 7(1), 210–218.
- [22] Talirongan, H., & al., et. (2025). Advanced image preprocessing strategies for mitigating challenges in optical character recognition of diverse textual images. *International Journal of Environmental Sciences*, 11(25s), 421–429.
- [23] Vaswani, A., & al., et. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [24] Vo, Q. N., Kim, S. H., Yang, H. J., & Lee, G. (2018). Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74, 568–586.
- [25] Wibowo, A. T., & Ardiansyah, F. (2020). Penerapan metode erosi dan dilasi pada citra digital untuk memperbaiki kualitas hasil OCR. *Jurnal Informatika Dan Komputasi*, 14(2), 112–120.
- [26] Wu, S., & Chang, F. (2025). A dual-engine fusion optical character recognition method for fast identification and key information extraction of drug labels. *Alexandria Engineering Journal*, 128, 1027–1036.
- [27] Zhang, Y., Liu, X., & Wang, Z. (2024). Advanced morphological image processing techniques for degraded document recognition. *Pattern Recognition Letters*, 178, 112–119.