

ABSTRAK

MUTIARA, 2026. Sistem Ekstraksi Teks Ter-Flag Plagiarisme Dari Laporan Plagiasi Berbasis Segmentasi Warna HSV Dan *Optical Character Recognition* (Dibimbing oleh **Rizki Yusliana Bakti, S.T., M.T** dan **Titin Wahyuni, S.Pd., M.T**).

Penelitian ini mengembangkan sistem ekstraksi teks ter-flag dari laporan *Turnitin* berbasis pengolahan citra digital, segmentasi warna HSV, dan OCR menggunakan Tesseract. Tujuan penelitian adalah mengubah *highlight Turnitin* yang bersifat visual menjadi data terstruktur yang dapat dianalisis secara kuantitatif. *Dataset* terdiri dari 106 dokumen PDF dengan total 3.057 halaman dan 7.530 *highlight* yang dianalisis. Metode yang digunakan mencakup konversi PDF ke citra, klasifikasi tipe halaman, segmentasi warna HSV, ekstraksi *bounding box*, *filter* geometri, serta ekstraksi teks menggunakan Tesseract (engine LSTM) dan verifikasi *text layer* PDF. Hasil penelitian menunjukkan bahwa *pipeline* berhasil mengidentifikasi 43,0 persen halaman yang mengandung *highlight*. Segmentasi HSV yang dikalibrasi secara empiris mampu mendeteksi delapan warna *match-group* secara konsisten. Proses OCR menghasilkan tingkat keterbacaan teks sebesar 76,7 persen pada kategori *good*. Sistem juga berhasil memisahkan *highlight body* dari *source-ID* label sehingga meningkatkan kualitas data ekstraksi. Selain itu, tahap klasifikasi tipe halaman menyaring halaman tanpa *highlight* sehingga proses komputasi menjadi efisien dan seluruh dataset dapat diproses dengan baik. Kesimpulan penelitian menunjukkan bahwa pendekatan berbasis pengolahan citra dan OCR mampu mengubah output *Turnitin* menjadi dataset terstruktur yang dapat digunakan untuk analisis lanjutan dan pengembangan model prediktif risiko plagiarisme.

Kata kunci: OCR, Plagiarisme, Segmentasi HSV.