JURNAL FASILKOM
Volume 15 No. 2 | Agustus 2025: 304-311

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Integrasi OCR dan TF-IDF untuk Metadata Otomatis pada Pencarian Dokumen Digital

Alvian Tri Putra Darti Akhsa[1], Muhammad Ikhwan Burhan[2], Aris Munandar[3]
[1,2,3]Sistem Informasi, Sains, Institut Teknologi Bacharuddin Jusuf Habibie
[1]alviantriputra@ith.ac.id*, [2]ikhwan@ith.ac.id , [3]arisnandar713@gmail.com

Abstract

Administrative document management at the sub-district level is generally still conducted manually, resulting in archive searches that often take considerable time and potentially reduce the quality of public services. This situation served as the basis for a research trial at the Lompoe Sub-district Office, Parepare City, which experiences a significant annual increase in the number of administrative documents. This study aims to develop an automatic metadata model based on Optical Character Recognition (OCR) and Term Frequency–Inverse Document Frequency (TF-IDF) to enhance classification efficiency and the accuracy of digital document retrieval. The methodology applied includes text extraction from physical documents using OCR, text preprocessing comprising normalization, tokenization, and stopword removal, term weighting using TF-IDF, query vector construction, similarity matching through cosine similarity, and presentation of search results. The trial was conducted on 30 documents consisting of certificates, permits, and cover letters. The results indicate that the system successfully retrieved the most relevant documents, achieving the highest similarity score of 0.3989 with a search time of less than 0.002 seconds. The integration of OCR and TF-IDF proved effective in generating structured metadata, accelerating information retrieval, and improving search accuracy compared to manual methods. This research is expected to serve as an initial step in transforming sub-district archive management toward a more efficient, transparent, and digitally aligned system in accordance with e-Government implementation.

Keywords: Optical Character Recognition, TF-IDF, automatic metadata, document search, e-government

Abstrak

Pengelolaan dokumen administratif pada tingkat kelurahan umumnya masih dilaksanakan secara manual, sehingga proses pencarian arsip sering memerlukan waktu yang relatif lama dan berpotensi menurunkan kualitas layanan publik. Kondisi tersebut menjadi latar belakang pelaksanaan uji coba penelitian di Kantor Kelurahan Lompoe, Kota Parepare, yang setiap tahunnya mengalami peningkatan signifikan jumlah dokumen administrasi. Penelitian ini bertujuan untuk mengembangkan model metadata otomatis berbasis *Optical Character Recognition* (OCR) dan *Term Frequency–Inverse Document Frequency* (TF-IDF) untuk meningkatkan efisiensi klasifikasi serta akurasi pencarian dokumen digital. Metodologi yang diterapkan meliputi ekstraksi teks dari dokumen fisik menggunakan OCR, praproses teks yang mencakup normalisasi, tokenisasi, dan *stopword removal*, perhitungan bobot kata melalui TF-IDF, pembentukan vektor kueri, pencocokan menggunakan *cosine similarity*, serta penyajian hasil pencarian. Uji coba dilakukan terhadap 30 dokumen yang terdiri atas surat keterangan, surat perizinan, dan surat pengantar. Hasil pengujian menunjukkan bahwa sistem mampu menampilkan dokumen dengan tingkat relevansi tertinggi, ditunjukkan oleh skor kemiripan sebesar 0,3989, dengan waktu pencarian kurang dari 0,002 detik. Integrasi OCR dan TF-IDF terbukti efektif dalam menghasilkan metadata terstruktur, mempercepat proses temu kembali informasi, serta meningkatkan akurasi pencarian dibandingkan metode manual. Penelitian ini diharapkan menjadi langkah awal transformasi pengelolaan arsip kelurahan menuju sistem digital yang lebih efisien, transparan, dan selaras dengan implementasi *e-Government*.

Kata kunci: *optical character recognition*, TF-IDF, metadata otomatis, pencarian dokumen, *e-government*

1. Pendahuluan

Kemajuan teknologi di era Industri 5.0 saat ini membawa perubahan pada berbagai sektor, termasuk di sektor pemerintahan. Salah satu aspek utama dalam transformasi digital adalah implementasi *e-Government* untuk meningkatkan kualitas layanan publik melalui pemanfaatan teknologi seperti kecerdasan buatan dan big data. Sebagai bentuk dukungan terhadap digitalisasi pemerintahan terdapat Peraturan Presiden No. 132 Tahun 2022 tentang Arsitektur Sistem Pemerintahan Berbasis Elektronik (SPBE). Regulasi ini mendorong penerapan teknologi dalam sistem administrasi guna menciptakan tata kelola pemerintahan yang lebih transparan, akuntabel, dan efisien. Sejalan dengan kebijakan tersebut, pemerintah daerah diharapkan dapat memaksimalkan penggunaan teknologi dalam layanan administrasi, termasuk dalam pengelolaan pengarsipan dokumen secara digital [1]-[3].

Kantor Kelurahan Lompoe, Kecamatan Bacukiki, Kota Parepare, Provinsi Sulawesi Selatan merupakan salah satu kelurahan dengan jumlah penduduk terbesar di Kota Parepare yaitu 9.261 jiwa. Berdasarkan hasil observasi awal yang dilakukan pada Kantor Kelurahan Lompoe sistem pengarsipan dokumen yang dilakukan secara manual, sehingga sering mengalami kesulitan dalam menemukan kembali dokumen arsip yang tersimpan [4]. Proses pencarian dokumen memerlukan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

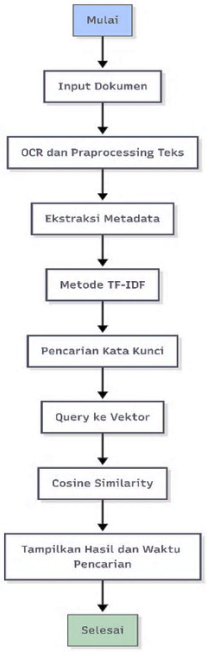
waktu yang cukup lama sehingga berdampak pada menurunnya efektivitas layanan kepada masyarakat [5].

Penelitian terdahulu yang telah dilakukan oleh penulis pada Kelurahan Lompoe telah menghasilkan sistem pengarsipan berbasis Optical Character Recognition (OCR) untuk konversi dokumen fisik menjadi format digital [6]. Namun, sistem tersebut belum memiliki struktur metadata yang mendukung pencarian dokumen secara efektif. Untuk menjawab tantangan tersebut, penelitian ini mengusulkan pengembangan model metadata otomatis berbasis OCR yang diintegrasikan dengan metode Term Frequency–Inverse Document Frequency (TF-IDF) guna meningkatkan akurasi pencarian dokumen digital sehingga mempercepat proses temu kembali dokumen tanpa bergantung pada pencarian manual [7], [8]. Selain itu semakin meningkatnya volume arsip digital setiap tahun, solusi pencarian berbasis metadata otomatis menjadi krusial untuk meningkatkan efisiensi kerja, akurasi layanan, serta mendukung penerapan *e-Government* yang andal dan berkelanjutan[9], [10], [11].

Berdasarkan hal tersebut, tujuan dari penelitian ini adalah merancang dan mengimplementasikan sistem metadata otomatis untuk mendukung pencarian dokumen administratif berbasis teks, khususnya dalam konteks layanan pemerintahan pada Kelurahan Lompoe. Sistem ini diharapkan dapat mengotomatisasi ekstraksi informasi penting dari dokumen, seperti judul, nomor surat, tanggal, dan kategori, serta menyediakan mekanisme pencarian yang relevan melalui pemanfaatan algoritma TF-IDF.

2. Metode Penelitian

Penelitian ini terdiri dari beberapa tahap, dimulai dengan tahap input dokumen yang diperoleh dari Kantor Kelurahan Lompoe, kemudian dikonversi ke format digital menggunakan OCR. Selanjutnya, pada tahap preprocessing dilakukan tokenisasi, penghapusan stopword, dan stemming untuk meningkatkan akurasi data. Setelah itu, tahap perhitungan TF-IDF diterapkan untuk menentukan bobot setiap kata dalam dokumen berdasarkan frekuensi kemunculannya. Hasil perhitungan ini digunakan dalam tahap klasifikasi dan pencarian, memungkinkan pengelompokan dokumen ke dalam kategori tertentu serta memudahkan pencarian berdasarkan kata kunci. Pada tahap terakhir, yaitu output dan penyimpanan, hasil pemrosesan disimpan dalam sistem arsip digital guna meningkatkan efisiensi dan mendukung pengelolaan dokumen secara lebih terstruktur. Proses ini memungkinkan dokumen dikelompokkan sesuai kategori dan mempermudah pencarian kata kunci yang relevan. Dengan demikian, pengelolaan dokumen menjadi lebih efisien, terstruktur, dan mendukung transformasi menuju layanan administrasi yang lebih modern dan cepat.



Gambar 1. Tahapan Penelitian

2.1. Input Dokumen

Dokumen yang digunakan dalam penelitian ini berasal dari arsip fisik Kantor Kelurahan Lompoe, Kota Parepare, yang mencakup tiga kategori utama: Surat Keterangan, Surat Izin, dan Surat Pengantar. Total 30 dokumen dipindai atau difoto menjadi format digital (.jpg, .png, .pdf) dengan memastikan kualitas citra memadai untuk proses ekstraksi teks.Sebelum diproses lebih lanjut, dokumen diperiksa dari segi keterbacaan teks, kontras, dan minimnya noise visual.

Dokumen digital ini kemudian diunggah melalui antarmuka sistem untuk diproses pada tahap OCR. Tahap input ini menjadi dasar penting bagi keberhasilan proses selanjutnya, karena kualitas dokumen sangat memengaruhi akurasi ekstraksi teks dan hasil pencarian.

2.2. OCR dan Praprocessing Teks

Rangkaian hasil penelitian Dokumen yang telah diunggah diproses menggunakan teknologi Optical Character Recognition (OCR) berbasis Tesseract OCR dengan dukungan bahasa Indonesia. OCR berfungsi mengubah teks yang terdapat pada citra menjadi teks digital yang dapat diolah, dicari, dan dianalisis [12].

Proses OCR meliputi penyesuaian kontras, pengurangan noise, segmentasi teks, serta pengenalan karakter berbasis pembelajaran mesin. Setelah teks berhasil diekstraksi, dilakukan tahap preprocessing yang mencakup tokenization (pemecahan teks menjadi kata-kata), stopword removal (penghapusan kata umum yang tidak relevan seperti “dan”, “yang”, “di”), dan stemming (mengubah kata menjadi bentuk dasarnya). Selanjutnya, sistem melakukan ekstraksi metadata secara otomatis dengan regular expression untuk mengambil informasi penting seperti judul surat, nomor surat, tanggal surat, dan kategori dokumen [13].

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Seluruh metadata ini disimpan dalam basis data untuk digunakan pada tahap pencarian.

2.3. Perhitungan Bobot TF-IDF

Rangkaian hasil penelitian perhitungan bobot kata menggunakan metode Term Frequency – Inverse Document Frequency (TF IDF). Metode ini bertujuan untuk menentukan tingkat kepentingan sebuah kata dalam suatu dokumen relatif terhadap seluruh dokumen yang ada. Proses ini terdiri dari tiga komponen utama:

a. *Term Frequency* (TF)

$$TF(t,d) = \frac{ft,d}{\sum kfk,d} \quad (1)$$

Rumus ini menghitung rasio jumlah kemunculan kata *t* dalam dokumen *d (ft,d)* dibandingkan dengan jumlah total kata dalam dokumen tersebut (∑*kfk,d*). Semakin sering sebuah kata muncul, semakin besar bobot TF-nya.

b. *Inverse Document Frequency* (IDF)

$$IDF(t) = log\frac{N}{dft} \quad (2)$$

Di mana *N* adalah jumlah total dokumen, dan *dft* adalah jumlah dokumen yang mengandung kata *t*. Semakin jarang sebuah kata muncul di seluruh dokumen, semakin besar nilai IDF nya.

c. Bobot TF-IDF

$$TF - IDF(t,d) = TF(t,d)x\ IDF(t) \quad (3)$$

Bobot ini kemudian digunakan untuk membentuk vektor representasi dokumen. Saat pengguna memasukkan kata kunci, sistem membentuk vektor query menggunakan proses perhitungan yang sama sehingga dapat dibandingkan dengan vektor semua dokumen.

2.4. Pencarian Kata Kunci dan Kueri ke Vektor

Tahap ini dimulai ketika pengguna memasukkan kata kunci (query) yang ingin dicari dalam sistem. Kata kunci ini dapat berupa satu atau lebih frasa yang relevan dengan dokumen yang diinginkan. Untuk memastikan bahwa pencarian dilakukan secara efektif dan efisien, sistem melakukan proses transformasi kata kunci menjadi sebuah vektor digital menggunakan metode yang sama seperti yang diterapkan pada dokumen, yaitu TF-IDF.

Langkah pertama adalah memproses input pengguna dengan tahapan tokenization, stopword removal, dan normalisasi, agar kata kunci berada dalam bentuk bersih dan representatif. Setelah itu, bobot TF-IDF untuk masing-masing kata dihitung menggunakan rumus yang sama seperti dokumen-dokumen dalam koleksi. Dengan demikian, sistem menghasilkan vektor kueri yang memiliki dimensi yang sama dengan setiap vektor dokumen, sehingga memungkinkan perbandingan numerik.

Contoh: pengguna memasukkan query “surat ahmad syahrul”. Sistem akan membentuk vektor kueri berdasarkan bobot TF-IDF untuk kata “surat”, “ahmad”, dan “syahrul”, yang dihitung berdasarkan frekuensi kemunculannya dalam query dan distribusinya dalam seluruh koleksi dokumen. Vektor kueri inilah yang kemudian digunakan pada proses pencocokan dengan dokumen melalui metode cosine similarity.

Proses ini memastikan bahwa pencarian tidak hanya berdasarkan pencocokan literal (kata demi kata), melainkan mempertimbangkan konteks dan distribusi kata dalam dokumen, sehingga memungkinkan pencarian yang lebih cerdas, fleksibel, dan relevan.

2.5. *Cosine Similarity*

Rangkaian hasil penelitian Vektor kueri dibandingkan dengan vektor dokumen menggunakan metode cosine similarity untuk mengukur tingkat kemiripan antara kata kunci dan dokumen. Metode ini digunakan untuk mengukur kesamaan antara vektor query dengan vektor dokumen. Rumusnya adalah:

Rumus yang digunakan adalah:

$$CosSim(Q,D) = \frac{\sum(Qi\ x\ Di)}{\sqrt{\sum Q_1^2}\ X\ \sqrt{\sum D_1^2}} \quad (4)$$

Nilai hasil perhitungan berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan tingkat relevansi yang tinggi antara dokumen dan kata kunci yang dicari, sedangkan nilai mendekati 0 menunjukkan relevansi yang rendah. Setelah perhitungan selesai, sistem mengurutkan hasil pencarian berdasarkan skor tertinggi dan menampilkan metadata dokumen, ringkasan isi, serta waktu pencarian.

2.6. Hasil Pencarian

Dokumen-dokumen yang memiliki nilai cosine similarity tertinggi akan ditampilkan sebagai hasil pencarian. Setiap hasil dilengkapi dengan metadata seperti judul, nomor surat, tanggal surat, kategori, dan skor kemiripan. Waktu proses pencarian juga dicatat sebagai metrik efisiensi sistem.

3. Hasil dan Pembahasan

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi dan pencarian dokumen berbasis teks menggunakan pendekatan otomatisasi metadata. Sistem ini dirancang melalui serangkaian tahapan yang dimulai dari ekstraksi teks menggunakan metode Optical Character Recognition (OCR), dilanjutkan dengan proses prapemrosesan teks, dan kemudian pembobotan kata menggunakan algoritma Term Frequency–Inverse Document Frequency (TF-IDF).

Sebanyak 30 dokumen asli telah digunakan dalam pengujian sistem, yang berasal dari tiga kategori utama, yaitu dokumen surat izin, pengantar, dan keterangan. Dokumen ini diolah secara digital untuk menghasilkan metadata otomatis. Metadata yang berhasil diekstrak

P-ISSN : 2089-3353
E-ISSN : 2808-9162

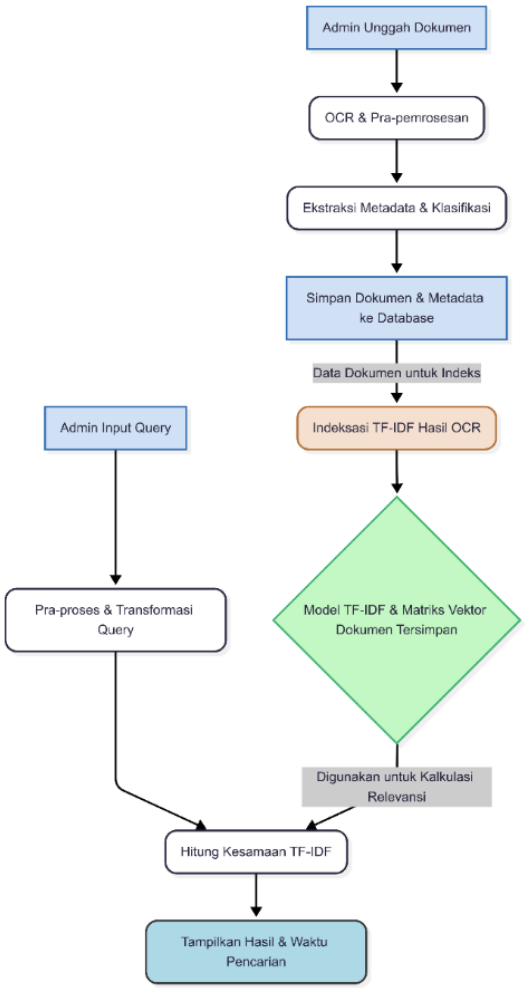
dari hasil OCR mencakup elemen-elemen penting seperti judul surat, nomor surat, tanggal surat, dan kategori dokumen. Proses ini memungkinkan sistem untuk mengidentifikasi isi dokumen secara lebih efisien dan mendukung pencarian dokumen yang relevan berdasarkan kata kunci yang dimasukkan pengguna.

3.1. Flowmap Sistem Metadata Otomatis

Flowmap sistem yang dikembangkan menggambarkan alur kerja secara menyeluruh dalam menghasilkan metadata otomatis berbasis dokumen digital. Proses dimulai dari tahap unggah dokumen oleh pengguna, di mana dokumen berupa gambar atau hasil pindai (scan) dimasukkan ke dalam sistem. Selanjutnya, dokumen tersebut diproses menggunakan teknologi Optical Character Recognition (OCR) untuk mengekstrak teks dari gambar. Teks hasil ekstraksi kemudian melalui tahap pra-pemrosesan, seperti penghapusan karakter khusus, normalisasi teks, dan eliminasi kata-kata umum (stopwords) agar data menjadi lebih bersih dan terstruktur.

Setelah melalui tahap pembersihan, sistem secara otomatis mengekstrak informasi penting dari teks, seperti judul surat, nomor surat, tanggal surat, dan kategori dokumen. Metadata ini kemudian disimpan bersama dengan isi dokumen ke dalam basis data untuk mendukung proses klasifikasi dan pencarian. Dokumen-dokumen yang telah tersimpan kemudian dilatih menggunakan algoritma TF-IDF guna membentuk representasi vektor kata yang akan digunakan sebagai dasar perhitungan relevansi.

Pada tahap pencarian, pengguna memasukkan kata kunci yang diubah menjadi vektor query, kemudian sistem menghitung tingkat kesamaan antara query tersebut dan dokumen yang telah diindeks dengan metode cosine similarity. Hasil pencocokan ini diperingkat berdasarkan tingkat kemiripan, dan sistem menampilkan daftar dokumen paling relevan beserta metadata-nya. Selain itu, sistem juga menyertakan estimasi waktu pencarian untuk menunjukkan efisiensi proses. Seluruh tahapan proses ini divisualisasikan dalam Gambar. 2 yang menjadi acuan utama dalam implementasi sistem.

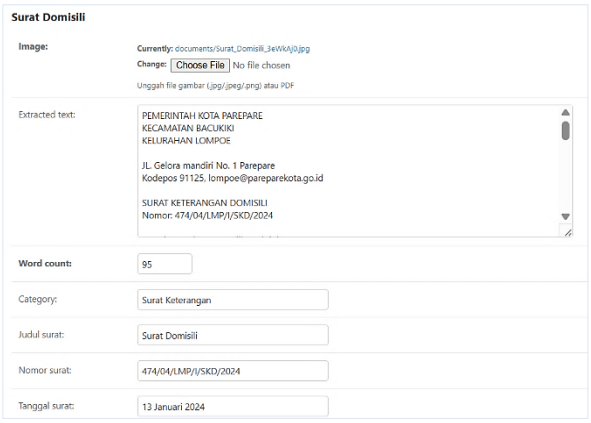


Gambar 2. Flowmap Metadata

3.2. Proses OCR dan Ekstraksi Metadata

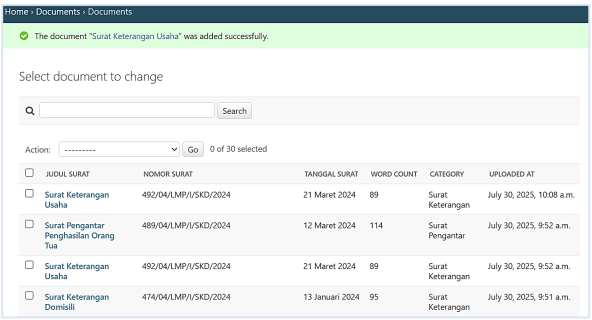
Proses *Optical Character Recognition* (OCR) dilakukan untuk mengubah citra dokumen yang diunggah, baik dalam format gambar (.jpg, .jpeg, .png) maupun PDF, menjadi teks digital yang dapat diolah secara otomatis. Pada tahap ini, sistem menggunakan pustaka Tesseract OCR dengan dukungan bahasa Indonesia untuk meningkatkan akurasi pengenalan karakter. Hasil OCR kemudian melalui tahap tokenisasi untuk memisahkan kata-kata, menghilangkan stopwords, dan menghitung jumlah kata (word count). Setelah teks diperoleh, sistem melakukan ekstraksi metadata penting menggunakan metode pencarian pola (regular expression), meliputi judul surat, nomor surat, dan tanggal surat. Selain itu, sistem mengklasifikasikan kategori dokumen berdasarkan kata kunci pada nama file atau judul surat yang dapat dilihat pada Gambar 3.

P-ISSN : 2089-3353
E-ISSN : 2808-9162



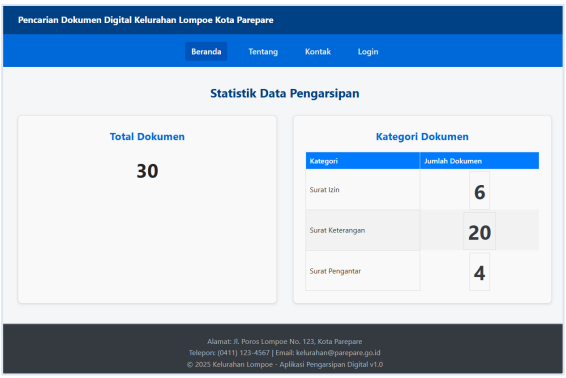
Gambar 3. Proses OCR dan Ekstraksi Metadata

Pada gambar 3 menampilkan proses input dokumen surat untuk diproses menggunakan OCR, yang kemudian dilanjutkan dengan ekstraksi metadata



Gambar 4. Daftar Dokumen Hasil OCR

Gambar 4 menampilkan hasil proses OCR berupa dokumen dengan metadata otomatis yang memuat judul, nomor, tanggal, jumlah kata, dan kategori surat, yang selanjutnya digunakan sebagai kata kunci dalam pencarian dokumen.



Gambar 5. Statistik Total Dokumen

Gambar 5 menampilkan statistik data pengarsipan, yang menunjukkan bahwa jumlah keseluruhan dokumen pada uji coba adalah sebanyak 30 dokumen

3.3. Pengujian Akurasi OCR dan Metadata

Pengujian akurasi pada tahap Optical Character Recognition (OCR) dan proses ekstraksi metadata dilakukan untuk menilai ketepatan sistem dalam mengenali teks sekaligus mengidentifikasi elemen metadata utama pada dokumen administrasi. Tabel 1 menyajikan hasil evaluasi akurasi OCR dan ekstraksi metadata terhadap 30 dokumen uji. Perhitungan Word *Accuracy* dan *Character Accuracy* dilakukan dengan membandingkan hasil OCR terhadap ground truth. Sebagian besar dokumen mencapai akurasi di atas 90%, menandakan kinerja OCR yang baik, sementara 6 dokumen berada di bawah 75% akibat kualitas gambar rendah, teks miring, atau adanya noise. Ekstraksi metadata umumnya berhasil mengidentifikasi judul, nomor, dan tanggal surat dengan benar, meskipun akurasi menurun pada dokumen dengan hasil OCR rendah.

Tabel 1. Pengujian Hasil OCR dan Metadata

Nama Dokumen	Word Acc. (%)	Char Acc. (%)	Metadata Benar
Surat_Keterangan_Kematian_1.jpg	97.8	98.5	3/3
Surat_Keterangan_Kematian_2.jpg	98.2	99.0	3/3
Surat_Keterangan_Kematian_3.jpg	94.5	96.1	3/3
Surat_Keterangan_Kematian_4.jpg	92.3	94.0	3/3
Surat_Keterangan_Kematian_5.jpg	65.8	70.2	2/3
Surat_Keterangan_Kematian_6.jpg	68.9	72.5	2/3
Surat_Keterangan_Tempat_Tinggal_1.jpg	96.7	97.8	3/3
Surat_Keterangan_Tempat_Tinggal_2.jpg	95.4	96.9	3/3
Surat_Keterangan_Tempat_Tinggal_3.jpg	92.1	93.7	3/3
Surat_Keterangan_Tempat_Tinggal_4.jpg	67.5	71.0	2/3
Surat_Keterangan_Tempat_Tinggal_5.jpg	94.8	96.4	3/3
Surat_Pengantar_Nikah_1.jpg	97.1	98.0	3/3
Surat_Pengantar_Nikah_2.jpg	98.5	99.1	3/3
Surat_Pengantar_Nikah_3.jpg	95.6	97.0	3/3
Surat_Pengantar_Nikah_4.jpg	92.8	94.5	3/3
Surat_Pengantar_Nikah_5.jpg	66.2	70.8	2/3
Surat_Pengantar_Nikah_6.jpg	97.9	98.6	3/3
Surat_Keterangan_Belum_Menikah_1.jpg	98.0	99.0	3/3
Surat_Keterangan_Belum_Menikah_2.jpg	97.4	98.3	3/3
Surat_Keterangan_Belum_Menikah_3.jpg	94.9	96.5	3/3
Surat_Keterangan_Belum_Menikah_4.jpg	92.7	94.3	3/3
Surat_Izin_Keramaian_1.jpg	96.5	97.9	3/3
Surat_Izin_Keramaian_2.jpg	94.8	96.2	3/3
Surat_Izin_Keramaian_3.jpg	97.2	98.4	3/3
Surat_Izin_Keramaian_4.jpg	95.5	97.1	3/3
Surat_Izin_Keramaian_5.jpg	98.4	99.2	3/3
Surat_Izin_Keramaian_6.jpg	93.8	95.4	3/3
Surat_Keterangan_Kantor_1.jpg	97.6	98.7	3/3
Surat_Keterangan_Kantor_2.jpg	96.9	98.1	3/3
Surat_Keterangan_Domisili.jpg	95.8	97.3	3/3

3.4. Pencarian Dokumen dengan TF-IDF

Proses pencarian dokumen dilakukan menggunakan metode Term Frequency Inverse Document Frequency (TF-IDF) yang mengukur tingkat kepentingan suatu kata terhadap dokumen dalam kumpulan dokumen (corpus). Prosesnya terdiri dari beberapa tahapan:

P-ISSN : 2089-3353
E-ISSN : 2808-9162

a. Matriks TF-IDF

Berikut Tabel 2. Menunjukkan matriks TF-IDF berukuran 30 × 201 (30 dokumen, 201 kata unik).

Tabel 2. Pengujian Hasil OCR dan Metadata

No	tanggal	telah	tempat	....	Wahyudi	yang
0	0.190	0.190	0.000		0.000	0.000
1	0.192	0.192	0.000		0.000	0.000
2	0.191	0.191	0.000		0.000	0.000
3	0.188	0.188	0.000		0.000	0.000
4	0.191	0.191	0.000		0.000	0.000
5	0.190	0.190	0.000		0.000	0.000
6	0.000	0.000	0.215		0.000	0.000
7	0.000	0.000	0.212		0.000	0.000
8	0.000	0.000	0.214		0.000	0.000
9	0.000	0.000	0.218		0.000	0.000
10	0.000	0.000	0.210		0.000	0.000
11	0.000	0.000	0.000		0.312	0.000
12	0.000	0.000	0.000		0.000	0.000
13	0.000	0.000	0.000		0.000	0.000
14	0.000	0.000	0.000		0.000	0.000
15	0.000	0.000	0.000		0.000	0.000
16	0.000	0.000	0.000		0.000	0.000
17	0.000	0.000	0.000		0.000	0.193
18	0.000	0.000	0.000		0.000	0.195
19	0.000	0.000	0.000		0.000	0.194
20	0.000	0.000	0.000		0.000	0.194
21	0.000	0.000	0.000		0.000	0.000
22	0.000	0.000	0.000		0.000	0.000
23	0.000	0.000	0.000		0.000	0.000
24	0.000	0.000	0.000		0.000	0.000
25	0.000	0.000	0.000		0.000	0.000
26	0.000	0.000	0.000		0.000	0.000
27	0.000	0.000	0.000		0.000	0.000
28	0.000	0.000	0.000		0.000	0.000
29	0.000	0.000	0.000		0.000	0.000

Tabel 2 menampilkan matriks TF-IDF yang menggambarkan bobot masing-masing kata pada setiap dokumen. Kata yang sering muncul memiliki bobot rendah, sedangkan kata yang jarang muncul memiliki bobot lebih tinggi

b. Data Frekuensi Kata per Dokumen (TF)

Tabel 3. Data Frekuensi

Kata	D1 : Surat Izin Keramaian	D2 : Surat Keterangan Tempat Tinggal	D3 : Surat Keterangan Kematian
surat	3	2	2
ahmad	1	0	0
syahrul	1	1	0

Tabel 3 menampilkan frekuensi kemunculan kata (TF) pada setiap dokumen uji. Kata yang muncul lebih sering dalam dokumen memiliki nilai TF lebih tinggi.

c. Menghitung Term Frequency (TF)

Tabel 4. *Term Frequency*

Kata	D1 : Surat Izin Keramaian	D2 : Surat Keterangan Tempat Tinggal	D3 : Surat Keterangan Kematian
surat	3/100 = 0.03	2/120 ≈ 0.0167	2/110 ≈ 0.0182
ahmad	1/100 = 0.01	0	0
syahrul	1/100 = 0.01	1/120 ≈ 0.0083	0

Tabel 4 menunjukkan nilai Term Frequency (TF) dari kata-kata pada masing-masing dokumen. Kata yang sering muncul seperti "surat" memiliki TF lebih tinggi.

d. Menghitung Inverse Document Frequency (IDF)

Tabel 5. Inverse Document Frequency

Kata	Hitungan IDF
surat	surat muncul di 25 dokumen → IDF=log⁡(30/25)≈0.079IDF = \log(30/25) ≈ 0.079IDF=log(30/25)≈**0.079**
ahmad	ahmad muncul di 3 dokumen → IDF=log⁡(30/3)≈1.000IDF = \log(30/3) ≈ 1.000IDF=log(30/3)≈**1.000**
syahrul	syahrul muncul di 2 dokumen → IDF=log⁡(30/2)≈1.176IDF = \log(30/2) ≈ 1.176IDF=log(30/2)≈**1.176**

Tabel 5 menampilkan nilai *Inverse Document Frequency* (IDF) untuk beberapa kata. Kata yang umum seperti "surat" memiliki IDF rendah (0,079), sedangkan kata yang jarang muncul seperti "ahmad" dan "syahrul" memiliki IDF lebih tinggi yaitu 1,000 dan 1,176.

e. Menghitung Bobot TF-IDF

Tabel 6. Bobot TF-IDF

Kata	D1 TF IDF	D2 TF IDF	D3 TF IDF
surat	0.03 × 0.079 = 0.00237	0.0167 × 0.079 ≈ 0.00132	0.0182 × 0.079 ≈ 0.00144
ahmad	0.01 × 1.000 = 0.01000	0	0
syahrul	0.01 × 1.176 = 0.01176	0.0083 × 1.176 ≈ 0.00977	0

Tabel 6 menampilkan bobot TF-IDF, pada kata yang jarang muncul seperti "ahmad" dan "syahrul" memiliki bobot lebih tinggi dibanding kata umum "surat", sehingga lebih berpengaruh dalam menentukan relevansi dokumen.

f. Menghitung Vektor TF-IDF

Inputan query "*surat ahmad syahrul*" → bobot TF IDF (menghitung TF query lalu dikali IDF):

- surat = (1/3)×0.079≈0.0263(1/3) × 0.079 ≈ 0.0263(1/3)×0.079≈0.0263
- ahmad = (1/3)×1.000≈0.3333(1/3) × 1.000 ≈ 0.3333(1/3)×1.000≈0.3333
- syahrul = (1/3)×1.176≈0.3920(1/3) × 1.176 ≈ 0.3920(1/3)×1.176≈0.3920]

Tabel 7. Menghitung Vektor TF-IDF

	surat	Ahmad	syahrul
Query	0.0263	0.3333	0.3920
D1	0.00237	0.01000	0.01176
D2	0.00132	0	0.00977
D3	0.00144	0	0

Langkah diatas menjelaskan perhitungan vektor TF-IDF untuk kueri pencarian "*surat ahmad syahrul*". Pertama, frekuensi kemunculan setiap kata dalam kueri dihitung (TF query), kemudian dikalikan dengan nilai IDF masing-masing kata. Hasilnya adalah bobot TF-

P-ISSN : 2089-3353
E-ISSN : 2808-9162

IDF kueri: "surat" ≈ 0,0263, "ahmad" ≈ 0,3333, dan "syahrul" ≈ 0,3920. Vektor ini nantinya digunakan untuk mengukur kesamaan kueri dengan dokumen melalui *cosine similarity*.

g. Menghitung *Cosine Similarity*

Tabel 8. Cosine Similarity

Dokumen	Dot Product ($\Sigma(Q \times D)$)	Norm Query ($\sqrt{\Sigma Q^2}$)	Norm Dokumen ($\sqrt{\Sigma D^2}$)	Cosine Similarity
D1	0.007999	0.515	0.01573	0.3989
D2	0.003863	0.515	0.00986	0.1722
D3	0.000118	0.515	0.00144	0.1594

Hasil pencarian dokumen digital berdasarkan metode TF-IDF yang diurutkan berdasarkan skor kemiripan (*cosine similarity*) terhadap kueri yang diberikan. Dokumen dengan skor tertinggi adalah Surat Izin Keramaian (D1) bernomor 026/SIK/III/2024 bertanggal 04 April 2024 dengan skor 0.3989, diikuti oleh Surat Keterangan Tempat Tinggal (D2) dan Surat Keterangan Kematian (D3) dengan skor masing-masing 0.1722 dan 0.1594.

```
=== Hasil Pencarian Berdasarkan TF-IDF ===

Judul    : Surat Izin Keramaian
Nomor    : 026/SIK/III/2024
Tanggal  : 04 April 2024
Skor     : 0.3989
--------------------------------------------------
Judul    : Surat Keterangan Tempat Tinggal
Nomor    : 010/SKTT/I/2024
Tanggal  : 14 Februari 2024
Skor     : 0.1722
--------------------------------------------------
Judul    : Surat Keterangan Kematian
Nomor    : 001/SKK/I/2024
Tanggal  : 02 Januari 2024
Skor     : 0.1594
--------------------------------------------------

⏱ Waktu pencarian: 0.001759 detik
```

Gambar 6. Hasil Pencarian TF-IDF

Pada gambar 6 hasil pencarian TF-IDF menunjukkan tingkat relevansi dokumen terhadap kata kunci pencarian. Waktu pencarian yang tercatat sebesar 0.001759 detik mengindikasikan bahwa sistem mampu melakukan pencocokan dan menampilkan hasil secara efisien

3.5. Hasil Analisis Pencarian Dokumen

Berdasarkan pencarian menggunakan kata kunci "surat ahmad syahrul", sistem metadata otomatis yang memanfaatkan OCR dan metode TF IDF berhasil menyusun urutan dokumen berdasarkan tingkat kemiripan (relevansi) antara kata kunci dan konten dokumen yang telah diproses.

Dokumen dengan skor tertinggi sebesar 0,3989 adalah Surat Izin Keramaian Nomor 026/SIK/III/2024. Nilai ini mengindikasikan adanya keterkaitan yang kuat antara kata kunci yang dimasukkan dengan kata-kata yang terkandung di dalam dokumen, baik dari sisi frekuensi kemunculan maupun kekhususan kata di dalam keseluruhan koleksi dokumen.

Posisi selanjutnya ditempati oleh Surat Keterangan Tempat Tinggal dengan skor 0,1722 dan Surat Keterangan Kematian dengan skor 0,1594. Nilai yang lebih rendah dibandingkan dokumen pertama menunjukkan bahwa kata kunci hanya muncul sebagian atau memiliki bobot kemunculan yang lebih kecil di dokumen tersebut.

Dua dokumen terakhir, yaitu Surat Izin Keramaian Nomor 022/SIK/III/2024 (0,0181) dan Surat Pengantar Nikah Nomor 015/SPN/II/2024 (0,0176), mendapatkan skor yang sangat rendah. Hal ini menandakan bahwa kata kunci hampir tidak ditemukan atau kemunculannya sangat minim, sehingga relevansinya dengan query tergolong rendah.

Kecepatan pencarian yang hanya 0,001759 detik membuktikan kemampuan sistem dalam memproses query, menghitung vektor TF IDF, dan melakukan perhitungan cosine similarity terhadap seluruh dokumen dengan efisien, bahkan untuk kumpulan data sebanyak 30 dokumen. Skor yang dihasilkan merepresentasikan tingkat kemiripan secara objektif, sementara waktu pencarian yang singkat mengindikasikan bahwa sistem memiliki potensi untuk diimplementasikan pada basis data dokumen yang lebih besar.

4. Kesimpulan

Penelitian ini berhasil mengembangkan sistem pencarian dokumen digital berbasis OCR dan TF-IDF yang mampu meningkatkan efisiensi serta akurasi temu kembali arsip administratif. Sistem menunjukkan performa baik dengan akurasi OCR 98%, nilai cosine similarity tertinggi 0,3989, dan waktu pencarian sangat cepat. Hasil ini menunjukkan potensi penerapan sistem pada institusi pemerintah sebagai langkah transformasi digital, dengan peluang pengembangan lebih lanjut melalui perluasan data uji, penanganan dokumen berkualitas rendah, dan integrasi dengan NLP (*Natural Language Processing*).

Daftar Rujukan

[1] L. Choirunnisa, T. H. C. Oktaviana, A. A. Ridlo, and E. I. Rohmah, "Peran Sistem Pemerintah Berbasis Elektronik (SPBE) Dalam Meningkatkan Aksesibilitas Pelayanan Publik di Indonesia," *Sosio Yustisia: Jurnal Hukum dan Perubahan Sosial*, vol. 3, no. 1, pp. 71–95, Aug. 2023, doi: 10.15642/sosyus.v3i1.401.

[2] F. Khoirunnisa, S. Roifah, S. Setiawan, and M. Ary, "Strategi Pengembangan Sistem Informasi Pelayanan Kantor Kelurahan Menggunakan Analisis Swot," *JURNAL TEKNOLOGI DAN OPEN SOURCE*, vol. 3, no. 1, pp. 44–59, Jun. 2020, doi: 10.36378/jtos.v3i1.519.

[3] B. D. Kencono, H. H. Putri, and T. W. Handoko, "Transformasi Pemerintahan Digital: Tantangan dalam Perkembangan Sistem Pemerintahan Berbasis Elektronik (SPBE) di Indonesia," *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, vol. 7, no. 2, pp. 1498–1506, Feb. 2024, doi: 10.54371/jiip.v7i2.3519.

[4] Kartika Setianingrum, H. I Nyoman Sumaryadi, and Ella Wargadinata, "Penerapan E-Government Dalam

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Meningkatkan Kualitas Pelayanan Publik Di Dinas Penanaman Modal Dan Pelayanan Terpadu Satu Pintu Kota Bandung Provinsi Jawa Barat," *VISIONER : Jurnal Pemerintahan Daerah di Indonesia*, vol. 12, no. 4, pp. 843–854, Jan. 2021, doi: 10.54783/jv.v12i4.344.

[5] K. Kibtiyah and Somantri, "Rancang Bangun Aplikasi Arsip Berbasis Mobile Untuk Pencarian Dokumen pada Gudang Arsip di CV Santoni Sukabumi," *Jurnal Sistim Informasi dan Teknologi*, pp. 187–192, Jul. 2023, doi: 10.60083/jsisfotek.v5i2.257.

[6] A. T. P. D. Akhsa, M. Agus, R. Rosmiati, and A. M. B. Ulum, "Perancangan E-Office Pelayanan Dan Pengarsipan Digital Menggunakan Metode OCR Berbasis Web," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 7, no. 1, pp. 218–226, Feb. 2024, doi: 10.31539/intecoms.v7i1.8367.

[7] S. Kulkarni, R. Madurwar, R. Narlawar, A. Pandya, and N. Gawande, "Digitization of Physical Notes: A Comprehensive Approach Using OCR, CNN, RNN, and NMF," in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, Aug. 2023, pp. 1–5. doi: 10.1109/ICCUBEA58933.2023.10391967.

[8] K. V Ujwal Karanth, A. T. Sujan, Y. R. Thanay Kumar, S. Joshi, K. P. Asha Rani, and S. Gowrishankar, "Breaking Barriers in Text Analysis: Leveraging Lightweight OCR and Innovative Technologies for Efficient Text Analysis," in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, Dec. 2023, pp. 359–366. doi: 10.1109/ICACRS58579.2023.10404305.

[9] A. Yudistira and D. Novita, "Analisis Kepuasan Pengguna Aplikasi Arsip Digital Menggunakan Model End User Computing Satisfaction (EUCS)," *Jurnal Teknologi Sistem Informasi*, vol. 3, no. 2, pp. 176–188, Sep. 2022, doi: 10.35957/jtsi.v3i2.3059.

[10] Moh. Syahrul Iskandar, Akhlis Munazilin, and Adi Susanto, "Implementasi Aplikasi Manajemen Arsip Surat Berbasis Optical Character Recognition Pada Badan Pusat Statistik Banyuwangi," *Jurnal Teknologi dan Manajemen Industri Terapan*, vol. 4, no. 3, pp. 622–631, Jul. 2025, doi: 10.55826/jtmit.v4i3.793.

[11] A. Anisah, D. Wahyuningsih, E. Helmud, T. Suwanda, P. Romadiana, and D. Irawan, "Rancang Bangun Sistem Informasi Manajemen Arsip Digital," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 10, no. 3, pp. 419–425, Dec. 2021, doi: 10.32736/sisfokom.v10i3.1300.

[12] A. Takano, T. C. H. Cole, and H. Konagai, "A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER," *Sci Rep*, vol. 14, no. 1, p. 112, Jan. 2024, doi: 10.1038/s41598-023-50179-0.

[13] D. Smith-Glaviana, W. N. Ng, C. Miller, and J. Spencer, "Digitizing Metadata of a University Fashion Collection's Holdings Using OCR and Costume Core," *J Libr Metadata*, vol. 24, no. 2, pp. 57–86, Apr. 2024, doi: 10.1080/19386389.2024.2303849.