JURNAL ILMIAH SISTEM INFORMASI (JUSI)
VOL 5, No. 1, Januari 2026, pp. 449 - 459
p-ISSN : 2803-1507 (media cetak) | e-ISSN : 2803-1531 (media online)
Link Publikasi: https://ejurnal.provisi.ac.id/index.php/JUISI/issue/view/169

Automatic Classification of Letters in the E-Archive System Using K-Nearest Neighbor

Beda Puspita Candra*[1], Achemad Fatkhur Mujaki[2]
[1] Sistem Informasi, Institut Teknologi Mojosari
Jl. Mojosari, Ds. Ngepeh, Loceret, Nganjuk, Jawa Timur, Indonesia
[2] Sistem Informasi, Universitas Pesantren Tinggi Darul 'Ulum
Jl. Komplek Pondok Pesantren Darul Ulum No. Tromol Pos 10, 61481, Jombang, Jawa Timur, Indonesia
e-mail: bedapuspita@itmnganjuk.ac.id*[1], vizaak3@gmail.com[2]

ARTICLE INFO

Article History:

Received: 12-11-2025
Revised: 15-12-2025
Accepted: 27-12-2025
Published: 21-01-2026

ABSTRACT

The research focuses on designing a web-based E-Archive application that automatically categorizes incoming letters using the K-Nearest Neighbors (KNN) classification method. The research was conducted at the Cultural Heritage Preservation Office of East Java, which continues to face challenges in managing archives manually, including difficulties in document retrieval, delays in information dissemination, and risks of data loss. The application was developed using PHP as the programming language and MySQL as the database engine, applying the Waterfall development model to ensure a structured process from requirements analysis through design, implementation, and testing. System testing used the Black-Box method to verify that all functions operate according to specifications. The classification process involves several stages, including text preprocessing (tokenization, stopword removal, and stemming), term weighting using Term Frequency–Inverse Document Frequency (TF-IDF), and cosine similarity for similarity measurement. Experimental results demonstrate an accuracy rate of 80% with an optimal K value of 14, indicating that the K-NN algorithm is effective in extracting and categorizing letter types. This E-Archive system is expected to improve efficiency in archive management, accelerate document search processes, and support administrative digitalization within government institutions. Furthermore, the research contributes to the practical application of text classification techniques in archival science. It provides a foundation for developing similar systems in other organizations that require modern and integrated document management solutions.

Keywords: *E-Archive, K-NN, TF-IDF, Digitalization, Archive Management.*

Abstrak

Penelitian ini berfokus pada perancangan aplikasi *E-Archive* berbasis web yang mampu melakukan kategorisasi otomatis terhadap surat masuk dengan menerapkan metode klasifikasi *K-Nearest Neighbor* (KNN). Penelitian dilakukan di Balai Pelestarian Cagar Budaya Jawa Timur, yang masih menghadapi berbagai kendala dalam pengelolaan arsip secara manual, seperti kesulitan pencarian dokumen,

DOI: https://doi.org/10.51903/336mw957

p-ISSN : 2803-1507 e-ISSN : 2803-1531

keterlambatan distribusi informasi, serta risiko kehilangan data. Aplikasi dikembangkan dengan memanfaatkan PHP sebagai bahasa pemrograman dan MySQL sebagai basis data, menggunakan model pengembangan Waterfall untuk memastikan proses yang terstruktur mulai dari analisis kebutuhan, desain, implementasi, hingga pengujian. Pengujian sistem dilakukan dengan metode Black Box untuk memverifikasi bahwa seluruh fungsi berjalan sesuai spesifikasi. Proses klasifikasi melibatkan beberapa tahapan, termasuk *text preprocessing* (tokenisasi, penghapusan *stopword*, dan *stemming*), pembobotan menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF), serta pengukuran kesamaan dokumen dengan *cosine similarity*. Hasil eksperimen menunjukkan tingkat akurasi sebesar 80% dengan nilai K optimal pada 14, yang menegaskan efektivitas algoritma KNN dalam mengekstraksi dan mengategorikan jenis surat. Sistem *E-Archive* ini diharapkan dapat meningkatkan efisiensi pengelolaan arsip, mempercepat proses pencarian dokumen, serta mendukung digitalisasi administrasi di instansi pemerintah. Selain itu, penelitian ini memberikan kontribusi praktis dalam penerapan teknik klasifikasi teks pada bidang kearsipan, sekaligus menjadi landasan bagi pengembangan sistem serupa di organisasi lain yang membutuhkan pengelolaan dokumen secara modern dan terintegrasi.

Kata Kunci: *E-Arsip, K-NN, TF-IDF, Digitalisasi, Menejemen Arsip.*

1. PENDAHULUAN

Digitalisasi arsip merupakan langkah strategis dalam meningkatkan efisiensi dan akurasi pengelolaan dokumen di lingkungan instansi pemerintah [1], [2], [3]. Salah satu lembaga publik yang bergerak di bidang pelestarian budaya menghadapi tantangan dalam klasifikasi jenis surat yang belum terotomatisasi, sehingga rentan terhadap kesalahan dan menyebabkan keterlambatan dalam pelayanan dalam proses administrasi. Pengawasan kearsipan diperlukan untuk menjamin pengelolaan arsip dinamis yang mampu mendukung kinerja secara cepat, tepat, dan sistematis [4]. Hasil penilaian internal atas manajemen arsip mengindikasikan bahwa proses penciptaan arsip telah berjalan optimal, sementara aspek pemanfaatan, perawatan, dan penyusutan masih membutuhkan peningkatan secara berkesinambungan. Hal ini menunjukkan urgensi penerapan sistem digital yang mampu mendukung klasifikasi dan pengelolaan arsip secara otomatis. Berbagai pendekatan telah dikembangkan untuk klasifikasi dokumen, termasuk metode berbasis citra dan teks. Metode klasifikasi berbasis *Convolutional Neural Network* (CNN) mampu mengenali pola visual dengan akurasi yang kompetitif, meskipun membutuhkan sumber daya komputasi tinggi [5]. Di sisi lain, metode berbasis teks seperti metode KNN dikenal sebagai solusi yang efisien dengan kebutuhan komputasi yang relatif rendah dan fleksibel untuk klasifikasi dokumen berbasis isi teks.

Selain itu, berbagai penelitian telah mengembangkan pendekatan klasifikasi dokumen, kinerja KNN dalam pengelompokan data teks menunjukkan hasil yang cukup baik, terutama ketika didukung oleh teknik preprocessing dan pembobotan yang tepat [6]. Hal ini memperkuat posisi KNN sebagai algoritma yang layak diterapkan dalam klasifikasi dokumen administratif, termasuk surat masuk dan keluar. Penelitian terbaru juga menunjukkan efektivitas KNN dalam klasifikasi data medis, yang memperkuat relevansi algoritma ini untuk dokumen administratif [7], [8] KNN juga telah digunakan dalam klasifikasi citra uang kertas dengan kombinasi Gabor Filter, yang menghasilkan akurasi hingga 97% [8] Selain itu, penerapan KNN dalam analisis sentimen ulasan produk elektronik di Tokopedia menunjukkan akurasi lebih dari 91% dengan dukungan teknik TF-IDF [9]

Lebih lanjut, dalam konteks pengelolaan arsip digital, sistem manajemen arsip berbasis cloud terbukti mampu meningkatkan efisiensi dan aksesibilitas dokumen [10]. Meskipun fokusnya pada penyimpanan, integrasi teknologi digital dalam sistem arsip memberikan dampak signifikan terhadap kualitas layanan informasi. Sejalan dengan itu, penelitian lain menunjukkan bahwa algoritma KNN dan pengembangannya (Improved KNN/IKNN) juga berhasil meningkatkan akurasi klasifikasi dalam analisis sentimen aplikasi digital, seperti pada studi terhadap aplikasi PaDi UMKM [11]. Hal ini menegaskan bahwa penerapan KNN dalam sistem berbasis web memiliki potensi luas, baik untuk pengelolaan arsip maupun layanan digital lainnya. Dengan mempertimbangkan temuan tersebut, studi ini difokuskan pada perancangan serta mengimplementasikan sistem E-Arsip berbasis web yang mampu melakukan klasifikasi jenis surat masuk secara otomatis menggunakan metode K-NN. Sistem ini diharapkan dapat meningkatkan efisiensi pengelolaan arsip dan mendukung digitalisasi administrasi di lingkungan instansi pemerintah.

2. TINJAUAN PUSTAKA

2.1. *Pengelolaan Arsip Digital*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Pengelolaan arsip digital bertujuan meningkatkan efisiensi, akurasi, dan aksesibilitas dokumen dalam instansi. Sistem informasi arsip berbasis web memungkinkan pencatatan, klasifikasi, dan pencarian dokumen secara terintegrasi [12]. Penerapan sistem ini dapat mengurangi duplikasi data, mempercepat proses pencatatan, dan mendukung transparansi administrasi publik. Klasifikasi dokumen merupakan proses pengelompokan berdasarkan karakteristik tertentu untuk mempermudah pencarian dan pengarsipan [13]. Dalam sistem arsip digital, klasifikasi berperan penting dalam pengelolaan arsip dinamis dan pengaturan akses dokumen secara sistematis.

2.2. *Text Mining dan Representasi Dokumen*

Teknik *Text Mining* dimanfaatkan dalam proses mengekstraksi informasi dari data teks melalui tahapan preprocessing seperti tokenisasi, stopword removal, dan stemming. Representasi dokumen dilakukan dengan metode TF-IDF, yang memberikan bobot pada kata-kata berdasarkan frekuensi dan relevansi dalam korpus [14]. Teknik ini memungkinkan pengukuran kemiripan antar dokumen menggunakan cosine similarity, sehingga dokumen dapat diklasifikasikan secara lebih akurat. [15] mengombinasikan pendekatan TF-IDF dengan algoritma KNN dalam mengklasifikasikan sentimen ulasan produk di *TikTok Shop*, dan berhasil mencapai akurasi 87,7%. Studi ini memperkuat bukti bahwa kombinasi TF-IDF dan KNN merupakan pendekatan yang andal dalam klasifikasi teks, sehingga mendukung penerapan serupa pada sistem E-Arsip berbasis web.

2.3. *Algoritma K-Nearest Neighbor (KNN)*

Metode K-Nearest Neighbor digunakan untuk mengklasifikasikan data dengan melihat tetangga terdekatnya, algoritma ini menggunakan metrik seperti cosine similarity atau Euclidean distance berfungsi menentukan kategori suatu data melalui analisis kedekatan dengan data tetangga terbanyak [16]. Ketepatan dalam memilih nilai k akan meningkatkan akurasi klasifikasi secara signifikan. KNN dikenal sederhana namun efektif, sehingga banyak digunakan dalam berbagai domain klasifikasi teks. KNN juga telah diimplementasikan dalam sistem monitoring hidroponik berbasis IoT, yang mampu mengklasifikasikan kondisi tanaman secara real-time dengan akurasi tinggi [17]

2.4. *Penerapan KNN dalam Penelitian Terdahulu dan Gap*

Sejumlah studi menunjukkan bahwa algoritma KNN memiliki performa kompetitif dalam berbagai domain klasifikasi. Dalam analisis sentimen ulasan aplikasi, KNN mencapai akurasi 80% dengan AUC 0.865 [18]. Pada konteks tanggap kebakaran, akurasi bahkan mencapai 98% [19]. sementara penelitian berbasis IoT untuk klasifikasi mutu telur ayam ras melaporkan akurasi 89–95% [20]. Di ranah bisnis, Mustaqim et al. [21] membuktikan efektivitas KNN dalam memprediksi kategori penjualan mobil dengan akurasi 71%, menegaskan fleksibilitas algoritma ini di luar domain teknis. Temuan ini menegaskan fleksibilitas KNN dalam berbagai bidang klasifikasi, sehingga relevan untuk diadaptasi dalam pengelolaan arsip digital [21].

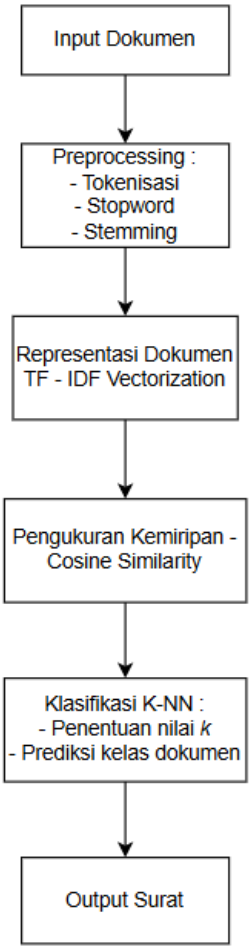
Selain itu, kombinasi KNN dengan representasi teks seperti TF-IDF juga terbukti relevan. Studi klasifikasi teks kesehatan mental menggunakan TF-IDF tanpa SMOTE menghasilkan akurasi 76.9%[14]. Penelitian lain menegaskan efektivitas KNN dalam domain medis maupun UMKM digital [10][13]. Sintesis dari berbagai temuan tersebut memperlihatkan bahwa KNN konsisten memberikan hasil yang baik, baik pada data numerik maupun teks, serta adaptif terhadap beragam konteks aplikasi. Namun demikian, mayoritas penelitian terdahulu masih berfokus pada kesehatan, UMKM, dan respon kebakaran. Belum banyak studi yang secara khusus mengintegrasikan KNN berbasis teks dengan sistem E-Arsip berbasis web untuk klasifikasi surat masuk di instansi pemerintah. Gap ini menegaskan urgensi penelitian yang dilakukan, dengan kontribusi pada pengembangan sistem arsip digital yang lebih efisien, adaptif, dan relevan bagi kebutuhan administrasi publik.

3. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk mengklasifikasikan dokumen surat masuk berdasarkan isi dan jenisnya menggunakan pendekatan pembobotan TF-IDF dan algoritma K-Nearest Neighbor (K-NN). Metode ini dipilih karena kesederhanaannya, efektivitasnya dalam klasifikasi teks, serta kemampuannya menangani data berukuran kecil hingga sedang. Pendekatan ini juga memungkinkan sistem untuk mengenali pola kemiripan antar dokumen secara matematis melalui pengukuran sudut antar vektor TF-IDF menggunakan cosine similarity. Dengan memanfaatkan nilai kemiripan tersebut, algoritma K-NN dapat menentukan kelas dokumen berdasarkan kedekatan dengan dokumen yang telah terlabel sebelumnya. Kombinasi antara

p-ISSN : 2803-1507 e-ISSN : 2803-1531

pembobotan kata dan klasifikasi berbasis tetangga terdekat ini memberikan solusi yang efisien dan terukur dalam pengelompokan dokumen surat masuk secara otomatis. Untuk memberikan gambaran menyeluruh, berikut diagram alur rangkaian metode yang dipakai tersaji pada Gambar 1.



Gambar 1. Diagram Alur Metode

3.1. *Input Dokumen*

Dokumen yang digunakan merupakan surat masuk dari instansi pemerintahan yang telah dikonversi ke format teks. Setiap dokumen dikategorikan berdasarkan jenis surat, seperti Lembaga, Instansi, Masyarakat, dan Pribadi.

3.2. *Preprocessing*

Proses ini bertujuan menyaring serta menyiapkan data agar siap digunakan sebelum dilakukan representasi numerik. Proses preprocessing meliputi:

1. **Tokenisasi**: Melakukan segmentasi teks ke dalam bentuk kata-kata.
2. **Stopword Removal**: Menghapus kata-kata umum yang tidak memiliki makna signifikan.
3. **Stemming**: Mengubah kata ke bentuk dasar.

Tahapan ini mengacu pada pendekatan yang digunakan oleh Pranata *et al.* , yang menerapkan preprocessing serupa dalam klasifikasi teks Twitter untuk pemilihan wakil presiden menggunakan TF-IDF dan SVM [22].

3.3. *Represntasi Dokumen*

Tahap berikutnya setelah preprocessing adalah merepresentasikan dokumen dalam bentuk vektor melalui teknik TF-IDF. Pada tahap ini TF-IDF digunakan untuk menekankan kata-kata penting dalam dokumen dan mengurangi pengaruh kata-kata umum. Rumus yang digunakan untuk mendapatkan hasil TF di tampilkan dalam Persamaan (1).

p-ISSN : 2803-1507 e-ISSN : 2803-1531

$$TF\ (d,t) = f(d,t) \quad (1)$$

Keterangan:

f(d,t) : Kemunculan kata t dalam dokumen d

Persamaan untuk menghitung nilai IDF ditunjukkan pada Persamaan (2).

$$IDF(t) = log\left({}^{n}/_{df(t)}\right) \quad (2)$$

Persamaan untuk menghitung TF-IDF divisualisasikan pada Rumus (3).

$$TFIDF = TF(d,t).IDF(t) \quad (3)$$

Nasrullah [23] menunjukkan bahwa TF-IDF mampu meningkatkan akurasi dalam deteksi kemiripan judul skripsi, terutama ketika dikombinasikan dengan pengukuran cosine similarity.

3.4. *Pengukuran Kemiripan (Cosine Similarity)*

Setelah dokumen direpresentasikan dalam bentuk vektor TF-IDF, tahap selanjutnya adalah mengukur kemiripan antar dokumen menggunakan cosine similarity. Teknik ini digunakan untuk mengukur sudut yang terbentuk antara dua vektor dalam ruang multidimensi. *Cosine similarity* menghasilkan nilai antara 0 sampai 1, dengan angka mendekati 1 menandakan kemiripan yang kuat. Pendekatan ini digunakan oleh Nasrullah [23], yang menunjukkan bahwa cosine similarity efektif dalam mendeteksi kemiripan judul skripsi secara otomatis berbasis TF-IDF.

3.5. *Klasifikasi dengan K-NN*

Langkah ini menerapkan algoritma *K-Nearest Neighbor* (KNN) sebagai metode klasifikasi untuk menentukan kelas dokumen berdasarkan kemiripan dengan dokumen lain. Prosesnya meliputi:

1. Penentuan nilai *k* secara eksperimental
2. Prediksi kelas dokumen berdasarkan tetangga terdekat.

Rumus algoritma KNN diperlihatkan pada Persamaan (4).

$$d = \sqrt{\sum_i^n (a_i - b_i)^2} \quad (4)$$

Keterangan:

d = jarak
a = data uji/testing
b = sampel data
i = variable data
n = dimensi data

Sukirman et al. [24] membuktikan bahwa algoritma *K-Nearest Neighbor* mampu menghasilkan performa klasifikasi yang cukup unggul dalam analisis sentimen *TikTok*, terutama ketika dikombinasikan dengan TF-IDF sebagai metode pembobotan fitur.

3.6. *Output Kelas*

Hasil akhir dari proses klasifikasi adalah keluaran berupa jenis surat dari dokumen yang dianalisis, yaitu surat Lembaga, Instansi, Masyarakat, dan Pribadi.

4. HASIL DAN PEMBAHASAN

Bab ini menampilkan hasil penerapan sistem yang telah dikembangkan yaitu sistem klasifikasi dokumen surat masuk, yang mengombinasikan teknik TF-IDF dengan algoritma KNN sebagai dasar analisis. Pengujian dilakukan terhadap 60 data surat, terdiri dari 50 data training dan 10 data testing. Hasil klasifikasi dianalisis berdasarkan tahapan metode dan dibandingkan dengan studi terdahulu.

4.1. *Hasil Preprocessing*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Tahapan preprocessing dilakukan untuk menyederhanakan isi dokumen dan mengekstraksi kata-kata penting. Proses ini meliputi tokenisasi, penghapusan stopword, dan stemming menggunakan algoritma Nazief & Adriani. Hasilnya, kata-kata umum seperti "yang", "dengan", dan "kepada" dihilangkan, sementara kata-kata bermakna seperti balai, undang, dan permohonan dipertahankan. Pendekatan ini sejalan dengan penelitian Gabriella [25], yang menunjukkan bahwa preprocessing yang tepat dapat meningkatkan akurasi klasifikasi teks. Hindarto [26] juga menekankan bahwa kualitas preprocessing berdampak langsung pada performa algoritma klasifikasi.

4.2. *Representasi Dokumen (TF-IDF)*

Setelah preprocessing, dokumen direpresentasikan dalam bentuk vektor TF-IDF**.** Bobot dihitung berdasarkan frekuensi kata dalam dokumen dan keseluruhan korpus. Contoh:

1. Term *balai* memiliki DF = 6 dan IDF = 1
2. Term *bapak* memiliki DF = 1 dan IDF = 1.778151
3. Bobot akhir dihitung sebagai TF \times IDF

Gabriella [25] menggunakan TF-IDF untuk mengekstraksi fitur penting dari dokumen pendaftaran siswa, dan menyatakan bahwa metode ini efektif untuk menangkap kata-kata kunci yang relevan dalam klasifikasi.

4.3. *Pengukuran Kemiripan (Cosine Similarity)*

Metode cosine similarity dipakai untuk mengukur kesamaan dokumen uji terhadap dokumen latih, di mana skor 0–1 menunjukkan tingkat kemiripan, dan nilai tinggi menandakan kemiripan kuat. Contoh hasil:

1. Dokumen T3 memiliki kemiripan tertinggi dengan D26, D27, D37, dan D4
2. Nilai cosine similarity tertinggi: 1.74699.

Wijaya [27] menerapkan cosine similarity dalam analisis sentimen Masyarakat terhadap pelayanan digital, dan menunjukkan bahwa pendekatan ini efektif dalam mengelompokkan opini berdasarkan kemiripan kata.

4.4. *Hasil Klasifikasi dengan KNN*

Klasifikasi dilakukan menggunakan algoritma KNN dengan nilai k = 14. Sistem memilih 14 dokumen terdekat berdasarkan skor cosine similarity, lalu menentukan kelas berdasarkan mayoritas tetangga. Pendekatan ini memungkinkan sistem untuk melakukan klasifikasi berbasis kemiripan konten, bukan sekadar pencocokan kata kunci. Dengan mempertimbangkan 14 dokumen terdekat, sistem memperoleh konteks yang lebih luas dalam menentukan kategori dokumen uji, sehingga mengurangi risiko klasifikasi yang bias akibat outlier atau dokumen yang tidak representatif. Berikut hasil klasifikasi dapat tersaji pada Tabel 1.

Tabel 1. Klasifikasi 10 Data Testing

Dokumen	Data Testing	Kelas Prediksi
T1	Test 1	Pribadi
T2	Test 2	Instansi
T3	Test 3	Instansi
T4	Test 4	Instansi
T5	Test 5	Lembaga
T6	Test 6	Pribadi
T7	Test 7	Lembaga
T8	Test 8	Pribadi
T9	Test 9	Pribadi
T10	Test 10	Masyarakat

4.5. *Evaluasi Kinerja*

Hasil perhitungan sebelumnya menghasilkan kelas untuk setiap data uji, sehingga tahap berikutnya adalah menguji akurasi dari klasifikasi KNN. Hasil yang diperoleh ditunjukkan pada Tabel 2.

Tabel 2. Hasil klasifikasi Data Testing

Data Testing	Kelas (Kategori)	Hasil
Test 1	Pribadi	Sesuai
Test 2	Instansi	Sesuai
Test 3	Instansi	Sesuai
Test 4	Instansi	Sesuai
Test 5	Lembaga	Sesuai

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Test 6	Pribadi	Tidak Sesuai
Test 7	Lembaga	Sesuai
Test 8	Pribadi	Sesuai
Test 9	Pribadi	Tidak Sesuai
Test 10	Masyarakat	Sesuai

Tahap berikutnya adalah mengukur tingkat akurasi dari metode KNN dengan k = 14 dan diperoleh pada perhitungan yang ditunjukkan pada Persamaan (5).

$$Akurasi = \frac{8}{10} = 0.8 \text{ x } 100\% = 80\% \quad (5)$$

Nilai akurasi sistem klasifikasi surat masuk sebesar 80% dapat dikategorikan cukup tinggi untuk sistem berbasis teks sederhana. Hindarto [26] membandingkan akurasi KNN dengan Naive Bayes dan Decision Tree, dan menemukan bahwa KNN memberikan hasil yang stabil pada data aplikasi Android dengan akurasi di atas 75%. Asmaraloka et al. [28] melaporkan bahwa penerapan KNN dengan teknik ekstraksi fitur zoning pada dataset MNIST mampu mencapai akurasi lebih dari 90%. Perbandingan ini menunjukkan bahwa meskipun akurasi sistem klasifikasi surat masuk masih di bawah capaian pada domain citra, hasil tersebut tetap kompetitif mengingat kompleksitas bahasa alami dan variasi struktur dokumen administratif.

Namun, analisis lebih lanjut terhadap data uji menunjukkan adanya kasus salah klasifikasi. Kesalahan ini umumnya terjadi pada dokumen dengan struktur kalimat yang ambigu, penggunaan istilah administratif yang tumpang tindih antar kategori, serta variasi panjang teks yang memengaruhi representasi TF-IDF. Misalnya, surat dengan topik kepegawaian yang memuat istilah umum seperti “pengajuan” atau “permohonan” kadang terklasifikasi ke kategori lain karena kemiripan kosakata. Implikasi dari kesalahan klasifikasi ini cukup signifikan bagi sistem arsip digital. Surat yang ditempatkan pada kategori yang salah dapat menghambat proses pencarian dan pengelolaan dokumen, serta berpotensi menurunkan efisiensi administrasi publik. Oleh karena itu, akurasi bukan hanya indikator teknis, tetapi juga berdampak langsung pada kualitas layanan arsip.

Potensi peningkatan performa dapat dilakukan melalui beberapa strategi. Pertama, menambah jumlah data latih agar representasi kosakata lebih kaya dan distribusi antar kategori lebih seimbang. Kedua, mengintegrasikan metode pembanding seperti Naive Bayes atau SVM untuk mengevaluasi apakah model lain mampu mengurangi tingkat salah klasifikasi. Ketiga, penerapan teknik oversampling seperti SMOTE dapat membantu mengatasi ketidakseimbangan data antar kategori. Dengan langkah-langkah tersebut, sistem diharapkan mampu meningkatkan akurasi sekaligus menurunkan risiko kesalahan klasifikasi yang berdampak pada pengelolaan arsip digital.

4.6. *Implementasi Sistem*

4.6.1. *Tabel Surat Masuk*

Berikut adalah tampilan halaman tabel surat masuk yang digunakan menampilkan tabel daftar surat masuk yang telah ditambahkan oleh admin tersebut. Halaman ini digunakan untuk mengelola daftar surat masuk oleh kesekertariatan. Sebagaimana ditunjukkan dalam Gambar 2.

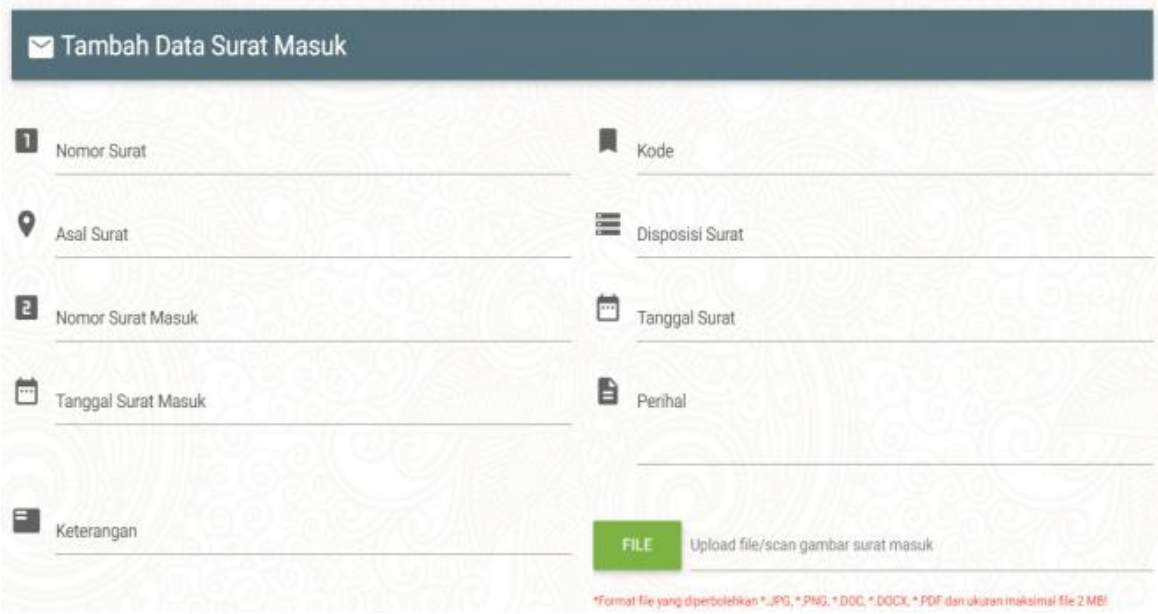
p-ISSN : 2803-1507 e-ISSN : 2803-1531



Gambar 2. Halaman Table Surat Masuk

4.6.2. *Tambah Data Surat*

Halaman Tambah Data surat masuk berisi kolom berisikan Nomor Surat, Asal Surat, Nomor Surat Masuk, Tanggal Surat Masuk, Kode Surat, Disposisi Surat, Tanggal Surat, Prihal, Keterangan , dan uplod file scan surat. Halaman ini digunakan untuk menambahkan data surat baru apabila memang terdapat surat yang masuk ke kesekertariatan. Visualisasi dapat diamati melalui Gambar 3.



Gambar 3. Halaman Tambah Data Surat

5. DISKUSI

Hasil penelitian menunjukkan bahwa kombinasi TF-IDF dan K-NN mampu mengklasifikasikan dokumen surat masuk secara otomatis dengan tingkat akurasi yang kompetitif. Proses preprocessing berperan penting

p-ISSN : 2803-1507 e-ISSN : 2803-1531

dalam meningkatkan kualitas vektor dokumen, sementara cosine similarity memberikan dasar matematis yang kuat dalam pemilihan tetangga terdekat. Dibandingkan dengan studi Gabriella [25] yang menerapkan TF-IDF untuk klasifikasi dokumen pendaftaran, dan Wijaya [27] yang menggunakan KNN untuk analisis sentimen, pendekatan ini terbukti fleksibel dan adaptif terhadap variasi isi dokumen. Sementara itu, akurasi sistem ini sebanding dengan hasil Hindarto [26], yang menunjukkan bahwa KNN tetap relevan dan kompetitif dibanding algoritma lain dalam klasifikasi dokumen digital.

6. KESIMPULAN DAN SARAN

6.1. *Kesimpulan*

Dari hasil penelitian, sistem klasifikasi dokumen surat masuk berhasil dikembangkan dengan metode TF-IDF sebagai pembobotan kata dan algoritma *K-Nearest Neighbor* (K-NN) sebagai teknik klasifikasi. Sistem ini mampu mengelompokkan dokumen ke dalam kategori Instansi, Lembaga, Masyarakat, dan Pribadi secara otomatis dengan tingkat akurasi mencapai 80%. Nilai k optimal yang digunakan adalah 14, dan proses preprocessing seperti tokenisasi, filtering, serta stemming turut meningkatkan kualitas hasil klasifikasi. Kelebihan sistem terletak pada kesederhanaan algoritma, kemudahan implementasi, dan fleksibilitas dalam menangani variasi isi dokumen. Namun, sistem masih memiliki keterbatasan, terutama sensitivitas terhadap kualitas data latih dan pemilihan nilai k, serta belum mampu menangani dokumen dengan struktur kalimat ambigu atau minim kata kunci. Selain itu, sistem belum dilengkapi dengan fitur validasi otomatis terhadap hasil klasifikasi.

6.2. *Saran*

Pengembangan selanjutnya yang disarankan meliputi integrasi metode klasifikasi lain seperti Naive Bayes atau Decision Tree untuk perbandingan performa, penambahan fitur visualisasi hasil klasifikasi, serta penerapan sistem ini dalam platform arsip digital instansi secara real-time.

DAFTAR PUSTAKA

[1] Anju Sitompul, Samsudin, and A. Priyamita, "Perancangan Sistem Pengarsipan Buku Tanah Berbasis Web (Studi Kasus:Badan Pertanahan Nasional Kab. Deli Serdang)," *Jurnal Ilmiah Sistem Informasi*, vol. 2, no. 1, pp. 20–29, Jan. 2023, doi: 10.51903/juisi.v2i1.547.

[2] N. Orinos, Q. Onola, and O. Ben Chistoff, "Zero-Shot Learning for Multilingual Document Classification in Low-Resource Languages," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 391–402, Dec. 2025, doi: 10.51903/jtie.v4i3.446.

[3] K. Khaeriyah, I. Amirullah, A. A. Rahman, N. Natsir, and N. F. Suyuthi, "Management of Entry and Outcomes of General Sub-Sections of Employment and Legal at the Education Office of South Sulawesi," *Journal of Management and Informatics*, vol. 1, no. 3, pp. 22–26, Dec. 2022, Accessed: Jan. 13, 2026. [Online]. Available: https://jmi.stekom.ac.id/index.php/jmi/article/view/12

[4] D. R. Aulianto, "Evaluasi Pengelolaan Arsip Dinamis pada Unit Pengolah di Sekretariat Utama Lembaga Ilmu Pengetahuan Indonesia," *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, vol. 8, no. 1, pp. 25–36, Jun. 2022, doi: 10.14710/lenpust.v8i1.44715.

[5] S. R. Suartika, I. W. S. E. Putra, A. Y. Wijaya, and R. Soelaiman, "Klasifikasi Citra Menggunakan Convolutional Neural Network (CNN) pada Caltech 101," *Jurnal Teknik ITS*, vol. 5, no. 1, pp. 1-76, 2016. doi: 10.12962/j23373539.v5i1.15696.

[6] L. A. Susanto, "Komparasi Model Support Vector Machine dan K-Nearest Neighbor pada Analisis Sentimen Aplikasi Polri Super APP," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, pp. 1180-1190, Apr. 2024, doi: 10.23960/jitet.v12i2.4152.

[7] T. A. Cahyaaty, H. Herlawati, A. Achmad, and H. Setiawan, "Model Prediksi Kondisi Kesehatan dari Data Medical Check-Up Menggunakan K-Nearest Neighbors dan Decision Tree," *JSRCS (Journal of Students Reasesarch in Computer Science*, vol. 6, no. 2, p. 135-148, Nov. 2025, doi: 10.31599/tvt7s936.

[8] A. Mulyani and N. P. Nurazizah, "Klasifikasi Keaslian Uang Kertas Menggunakan Algoritma K-Nearest Neighbor dan Metode Gabor Filter," *Jurnal Algoritma*, vol. 22, no. 2, pp. 1510-1520, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2375.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[9] N. F. Octavia and Berlilana, "Penerapan Algoritma K-Nearest Neighbor untuk Analisis Sentimen Ulasan Produk Elektronik pada Platform E-Commerce," *Jurnal Algoritma*, vol. 22, no. 2 pp. 2110-2121, Nov. 2025, doi: 10.33364/algoritma/v.22-2.3083.

[10] T. Kartawiyuda, N. Heryana, and A. A. Ridha, "Implementasi Owncloud pada Sistem Manajemen Arsip Digital Di Perpustakaan Universitas," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 2498-2505, Aug. 2024, doi: 10.23960/jitet.v12i3.4792.

[11] M. Febima, U. Solihin, L. Magdalena, M. Hatta, M. Asfi, and S. Christina, "Analisis Sentimen Aplikasi PaDi UMKM dengan Peningkatan Kinerja Algoritma KNN," *J. Digit.: Digital Inf. Technol.*, vol. 15, no. 2, pp. 70–82, 2025. doi: 10.51920/jd.v15i2.446.

[12] S. Naja, R. Akbar, and Ismail, "Sistem Informasi Pengelolaan Arsip Digital pada Kantor Dinas Pertanian Provinsi Aceh Berbasis Web," *Jurnal Sistem Komputer (SISKOM)*, vol. 4, no. 2, pp. 60–71, Aug. 2024, doi: 10.35870/siskom.v4i2.813.

[13] A. Husnul Hotimah, S. Sukaesih, and K. Kusnandar, "Implementasi Sistem Klasifikasi Keamanan dan Akses Arsip: Studi Kasus pada Records Center FISIP Universitas Padjadjaran," *ANUVA*, vol. 8, no. 4, pp. 613–626, 2024. doi: 10.14710/anuva.8.4.613-626.

[14] A. Firizkiansah, A. Muhammad, and I. R. Maulana, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," *J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 1, pp. 29–36, May 2025. [Online]. Available: https://ojs.sains.ac.id/index.php/jikomti/article/view/97

[15] M. Murtiwiyati, R. F. Arini, L. Safitri, S. Widianti, and I. Setiadi, "Implementasi Algoritma K-Nearest Neighbors untuk Klasifikasi Sentimen pada Ulasan Pakaian Anak di Platform TikTok Shop," *Jurnal Minfo Polgan*, vol. 14, no. 2, pp. 2921–2928, Nov. 2025, doi: 10.33395/jmp.v14i2.15538.

[16] R. B. Trianto and A. S. Nugroho, "Peringkasan Dokumen Teks Bilingual Sebagai Reduksi Fitur untuk Klasifikasi Menggunakan Algoritma K-NN," *LogicLink: Journal of Artificial Intelligence and Multimedia in Informatics*, vol. 1, no. 1, pp. 37–49, Jun. 2024. doi: 10.28918/logiclink.v1i1.7801.

[17] A. S. Wibisono, N. Ningsih, and A. D. Ramadhani, "Implementasi Sistem Pemantauan Tanaman Hidroponik Berbasis IoT Menggunakan Metode KNN," *Proc. Seminar Nas. Terapan Riset Inovatif (SENTRINOV)*, vol. 11, no. 1, pp. 300–307, 2025. [Online]. Available: https://proceeding.isas.or.id/index.php/sentrinov/article/view/1711

[18] D. Pratmanto, R. Wijianto, A. Widayanto, and Ubaidilah, "Komparasi K-Nearest Neighbors (KNN) dan Naive Bayes pada Klasifikasi Sentimen Ulasan Aplikasi Tokopedia di Google Play Store," *Informatics and Computer Engineering Journal*, vol. 5, no. 2, pp. 75–80, Aug. 2025. [Online]. Available: https://jurnal.bsi.ac.id/index.php/ijec/article/view/8939

[19] D. M. Jeffri and A. Witanti, "Klasifikasi Waktu Tanggap Kebakaran oleh Pemadam Kebakaran di Kabupaten Sleman Menggunakan Metode k-Nearest Neighbor," *Jutisi: J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 14, no. 2, pp. 1227–1236, Aug. 2025. doi: 10.35889/jutisi.v14i2.2966.

[20] S. Delfi Mardiana Putri, M. Berlian Sari, and M. Kudus Sussalam, "Rancang Bangun Sistem Kualitas Telur Ayam Ras Berbasis Internet of Things untuk Transformasi Digital Peternakan," *BIOCHEPHY: Journal of Science Education*, vol. 5, no. 2, pp. 991–1001, 2025, doi: 10.52562/biochephy.v5i2.1773.

[21] K. Mustaqim, F. Amalia Maresti, S. Salwa Azzahra, and S. Paringga Pangestu Ningsih, "Analisis Faktor yang Mempengaruhi Penjualan Mobil Berdasarkan Model k-Nearest Neighbour," *Data Sciences Indonesia (DSI)*, vol. 5, no. 2, pp. 10–19, Dec. 2025. doi: 10.47709/dsi.v5i2.5667.

[22] R. Albin Pranata, N. Azmi Verdikha, U. Muhammdiyah Kalimantan Timur, and J. H. Ir Juanda, "Metode Pembobotan Tf-Idf untuk Klasifikasi Teks Quick Count Pemilihan Wakil Presiden Indonesia 2024 Pada X Twitter Dengan Metode SVM," vol. 18, no. 2, p. 126, Aug. 2024, doi: 10.47111/jti.

[23] A. H. Nasrullah, "Integrasi tf-idf dan Algoritma Cosine Similarity untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer Unisan Gorontalo)," *Information Technology Education Journal*, vol. 3, no. 3, pp. 113–118, Dec. 2024. [Online]. Available: http://journal.unm.ac.id/index.php/intec/article/view/5810

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[24] Sukirman, Sajiah, N. Putri Husain, A. Febriana Syam, and R. Mustikosari, "Analisis Sentimen Ulasan Pengguna Tiktok pada Google Play Store Berbasis tf-idf dan Support Vector Machine," *J. Syst. Comput. Eng.*, vol. 5, no. 1, pp. 91–102, 2024. doi: 10.61628/jsce.v5i1.1105.

[25] Y. A. Putri Gabriella, "Optimasi Penerimaan Siswa Baru Dengan Penerapan Algortima Text Mining Dan Tf-Idf," *Journal of Computing and Informatics Research*, vol. 2, no. 3, pp. 110–117, Jul. 2023, doi: 10.47065/comforch.v2i3.941.

[26] D. Hindarto, R. E. Indrajit, and E. Dazki, "Perbandingan Kinerja Akurasi K-NN, NB dan DT pada APK Android," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 1, pp. 486–503, 2022. doi: 10.35957/jatisi.v9i1.1542.

[27] D. Wijaya and S. Sutedi, "Analisis Sentimen Opini Masyarakat Terhadap Penerapan Pelayanan Digital di Desa Margo Mulyo Menggunakan Algoritma k-Nearest Neighbor," *J. Ilmu Komput., Sist. Inf., Tek. Inf.*, vol. 3, no. 1, pp. 13–26, Mar. 2024. [Online]. Available: https://jurnal.akommedia.net/index.php/jilkomsiti/article/view/45

[28] K. Nisa, M. A. Hermansyah, A. M. Asmaraloka, F. H. D. Saputra, and A. Setiawan, "Implementation of the K-Nearest Neighbor (KNN) Algorithm in Handwritten Digit Pattern Recognition Using the Zoning Method," *Jurnal Ilmiah Sistem Informasi*, vol. 4, no. 2, pp. 175–185, May 2025, doi: 10.51903/kf6s5f56.