Analisis Klasifikasi Surat BPS Sragen Menggunakan TF-IDF dan Algoritma Multinomial Naïve Bayes pada E-Arsip

Fitria Ayu Novitasari[1*], Wiji Lestari[2], Pramono[3]

[1]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[1*]220103124@mhs.udb.ac.id

[2]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[2]wiji_lestari@udb.ac.id

[3]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[3]pramono@udb.ac.id

Abstrak— **Pengelolaan arsip surat pada Badan Pusat Statistik (BPS) Kabupaten Sragen masih dilakukan secara manual sehingga proses pengelompokan dan pencarian surat memerlukan waktu yang relatif lama serta berpotensi menimbulkan kesalahan klasifikasi. Penelitian ini bertujuan menerapkan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) dan algoritma Multinomial Naive Bayes untuk melakukan klasifikasi otomatis jenis surat pada sistem e-arsip berbasis web. Tahapan penelitian meliputi pengumpulan dataset surat, preprocessing teks, pembobotan kata menggunakan TF-IDF, serta proses klasifikasi menggunakan Multinomial Naive Bayes. Sistem dikembangkan menggunakan bahasa pemrograman Python dengan *framework* Flask. Dataset yang digunakan terdiri atas tiga kategori surat, yaitu pengumuman, undangan, dan permohonan. Hasil pengujian menunjukkan bahwa sistem mampu mengklasifikasikan jenis surat secara otomatis dengan tingkat akurasi sebesar 85,71%. Hasil tersebut menunjukkan bahwa kombinasi metode TF-IDF dan algoritma Multinomial Naive Bayes cukup efektif dalam meningkatkan efisiensi pengelolaan arsip surat digital. Dengan demikian, sistem yang dikembangkan dapat membantu proses pengelompokan dan pencarian arsip surat secara lebih cepat, tepat, dan terstruktur pada BPS Kabupaten Sragen.**
Kata kunci— *Naive Bayes*, TF-IDF, Klasifikasi Surat, E-Arsip, *Text Mining.*

Abstract— **Mail archive management at the Central Statistics Agency (BPS) of Sragen Regency is still carried out manually, causing the grouping and retrieval process to take a relatively long time and increasing the risk of classification errors. This study aims to implement the Term Frequency–Inverse Document Frequency (TF-IDF) method and the Multinomial Naive Bayes algorithm for automatic classification of mail types in a web-based e-archive system. The research stages include mail dataset collection, text preprocessing, term weighting using TF-IDF, and classification using the Multinomial Naive Bayes algorithm. The system was developed using the Python programming language and the Flask framework. The dataset consists of three categories of letters, namely announcements, invitations, and requests. The testing results show that the system is capable of automatically classifying mail types with an accuracy rate of 85.71%. These results indicate that the combination of the TF-IDF method and the Multinomial Naive Bayes algorithm is sufficiently effective in improving the efficiency of digital mail archive management. Therefore, the developed system can assist in the grouping and retrieval of mail archives more quickly, accurately, and systematically at BPS Sragen Regency.**
Keywords— Naive Bayes, TF-IDF, Mail Classification, E-Archive, Text Mining.

I. PENDAHULUAN

Perkembangan teknologi informasi mendorong berbagai instansi untuk melakukan transformasi digital dalam pengelolaan data dan dokumen. Salah satu bentuk transformasi tersebut adalah penerapan sistem electronic archive (e-arsip) yang digunakan untuk menyimpan, mengelola, dan mencari dokumen secara digital. Penggunaan sistem e-arsip mampu meningkatkan efisiensi kerja, mempercepat proses pencarian dokumen, serta meminimalkan risiko kehilangan arsip dibandingkan pengelolaan arsip secara konvensional [1].

Badan Pusat Statistik (BPS) Kabupaten Sragen merupakan instansi pemerintah yang memiliki aktivitas administrasi surat menyurat cukup tinggi, baik surat masuk maupun surat keluar. Pengelolaan jenis surat yang masih dilakukan secara manual menyebabkan proses pengelompokan arsip membutuhkan waktu yang cukup lama dan berpotensi menimbulkan kesalahan klasifikasi. Selain itu, banyaknya dokumen surat yang tersimpan juga menyulitkan proses pencarian kembali arsip ketika dibutuhkan. Oleh karena itu, diperlukan suatu sistem yang mampu melakukan klasifikasi jenis surat secara

otomatis agar pengelolaan arsip menjadi lebih efektif dan efisien [2].

Klasifikasi dokumen teks merupakan salah satu implementasi *text mining* yang banyak digunakan untuk mengelompokkan dokumen berdasarkan kategori tertentu. Salah satu algoritma yang sering digunakan dalam klasifikasi teks adalah Naive Bayes karena memiliki proses komputasi yang sederhana, cepat, dan mampu memberikan tingkat akurasi yang baik pada dokumen teks [3]. Dalam proses klasifikasi teks, diperlukan metode pembobotan kata untuk menentukan tingkat kepentingan suatu kata pada dokumen. Metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk memberikan bobot kata berdasarkan frekuensi kemunculan kata dalam dokumen dan keseluruhan dataset sehingga dapat meningkatkan performa klasifikasi [4].

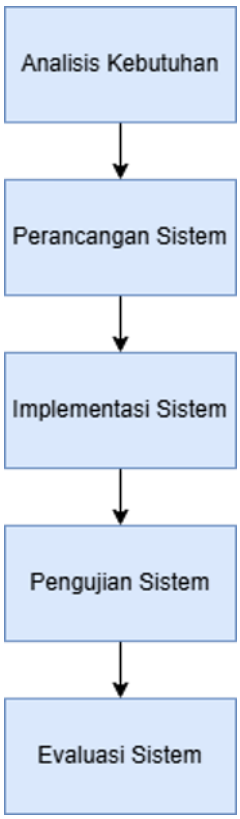
Beberapa penelitian terdahulu menunjukkan bahwa kombinasi metode TF-IDF dan algoritma Multinomial Naïve Bayes efektif digunakan dalam proses klasifikasi teks. Hal ini sejalan dengan penelitian yang dilakukan oleh Saputra dkk. yang menunjukkan bahwa penerapan pembobotan TF-IDF bersama Multinomial Naïve Bayes mampu menghasilkan performa klasifikasi yang baik pada dokumen teks, khususnya dalam meningkatkan hasil evaluasi model [4]. Penelitian oleh Manikome dan Maramis menunjukkan bahwa penerapan algoritma K-Means pada sistem arsip surat masuk dan surat keluar berbasis web di instansi pemerintahan dapat membantu dalam mengelompokkan dokumen secara lebih terstruktur dan efisien sesuai dengan karakteristik datanya [5].

Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma Naive Bayes dengan metode TF-IDF untuk melakukan klasifikasi otomatis jenis surat pada sistem e-arsip BPS Sragen. Sistem dikembangkan berbasis web menggunakan bahasa pemrograman Python dan *framework* Flask. Dataset yang digunakan terdiri dari beberapa kategori surat, yaitu pengumuman, undangan, dan permohonan. Penelitian ini diharapkan dapat membantu proses pengelolaan arsip surat secara digital sehingga proses klasifikasi surat dapat dilakukan dengan lebih cepat, tepat, dan efisien.

II. METODOLOGI PENELITIAN

A. Metode Penelitian

Pengembangan sistem pada penelitian ini menggunakan metode Waterfall karena memiliki tahapan yang terstruktur dan sistematis dalam proses pengembangan perangkat lunak. Metode ini terdiri dari beberapa tahapan, yaitu analisis kebutuhan, perancangan sistem, implementasi, pengujian, dan pemeliharaan sistem. Alur tahapan metode Waterfall pada penelitian ini ditunjukkan pada Gambar 1.

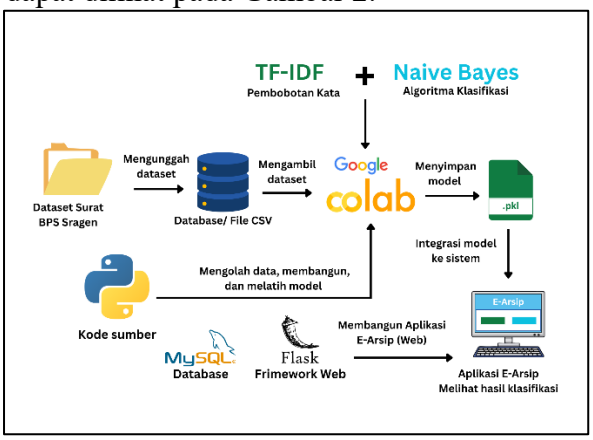


Gambar 1. Metode waterfall

B. Tahapan Penelitian

Tahapan penelitian ini dimulai dari pengumpulan dataset surat yang diperoleh dari BPS Kabupaten Sragen. Dataset kemudian melalui proses preprocessing untuk membersihkan data teks sebelum dilakukan pembobotan kata menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Hasil pembobotan digunakan sebagai

fitur masukan pada algoritma Naive Bayes untuk melakukan klasifikasi jenis surat secara otomatis [6]. Model yang telah dibangun selanjutnya diintegrasikan ke dalam sistem e-arsip berbasis web. Alur tahapan penelitian dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur model

C. Dataset dan Preprocessing Data

Dataset terdiri dari beberapa jenis surat, yaitu surat pengumuman, surat undangan, dan surat permohonan yang digunakan sebagai data latih (*training data*) dan data uji (*testing data*) dalam proses klasifikasi otomatis. Pengelolaan dokumen dilakukan melalui sistem electronic archive (e-arsip), yaitu sistem pengelolaan dokumen digital yang digunakan untuk menyimpan, mengelola, dan mencari arsip secara lebih efektif dan efisien. Penerapan e-arsip dapat meningkatkan keamanan data, mempercepat proses pencarian dokumen, serta mengurangi risiko kehilangan arsip [7]. Selain itu, sistem berbasis web memungkinkan pengelolaan arsip menjadi lebih terstruktur dan mudah diakses sesuai kebutuhan pengguna [8].

Dalam penelitian ini, proses pengolahan data dilakukan menggunakan pendekatan *text mining*, yaitu teknik untuk mengubah data teks yang tidak terstruktur menjadi informasi yang lebih terstruktur sehingga dapat dianalisis [9]. Penerapan *text mining* bertujuan untuk membantu sistem mengenali pola kata pada setiap kategori surat sehingga proses pengelompokan dokumen dapat dilakukan secara lebih cepat dan akurat [10]. Dokumen surat yang digunakan berbentuk teks hasil input manual maupun dokumen digital yang telah diproses pada sistem e-arsip. Total dataset yang digunakan sebanyak 105 data surat yang terdiri atas 39 surat pengumuman, 35 surat undangan, dan 31 surat permohonan. Deskripsi dataset penelitian dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi Dataset Penelitian

Variabel	Keterangan	Kategori Data
Jenis Surat	Pengumuman, Permohonan, Undangan	Output
Isi Surat	Teks dokumen surat	Input
Jumlah Data	105 dokumen	Dataset
Format Data	Dokumen Teks	Text
Sumber Data	BPS Kabupaten Sragen	Sekunder

Sebelum dilakukan proses klasifikasi, data teks terlebih dahulu melalui tahapan *preprocessing* untuk membersihkan dan menyiapkan data agar lebih optimal. Tahapan ini meliputi *case folding*, *cleaning*, *tokenizing*, dan *filtering*. Proses *case folding* dilakukan dengan mengubah seluruh huruf menjadi huruf kecil, sedangkan *cleaning* digunakan untuk menghapus angka, tanda baca, dan karakter khusus. Selanjutnya, tahap *tokenizing* dilakukan dengan memecah kalimat menjadi kumpulan kata, kemudian dilanjutkan dengan *filtering* untuk menghapus kata-kata yang tidak memiliki makna penting. Tahapan *preprocessing* bertujuan untuk meningkatkan kualitas data teks sehingga proses klasifikasi dapat dilakukan secara lebih optimal [11]. Setelah proses tersebut selesai, dataset dibagi menjadi data latih dan data uji dengan perbandingan 80% : 20% menggunakan metode *train-test split*.

D. Metode Klasifikasi TF-IDF dan Naive Bayes

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan kata yang digunakan untuk menentukan tingkat kepentingan suatu kata dalam dokumen. Metode ini bekerja dengan menghitung frekuensi kemunculan kata pada dokumen dan membandingkannya dengan jumlah kemunculan kata pada seluruh dokumen dalam dataset [12].

$$IDF(t) = log\frac{N}{df(t)} \quad (1)$$

Keterangan:
N= Jumlah dokumen.

Df(t)= Jumlah dokumen yang mengandung term.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (2)$$

Keterangan:
TF-IDF(t,d) = nilai bobot kata *term* t pada dokumen d.
TF(t,d) = jumlah kemunculan kata (*term frequency*) $t$pada dokumen d.
IDF(t) = nilai *inverse document frequency* dari kata t , digunakan untuk mengukur tingkat pentingnya kata dalam seluruh dokumen.
t = *term* atau kata yang dianalisis.
d = dokumen yang diproses.

Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes untuk menentukan kategori suatu data. Algoritma ini memiliki keunggulan dalam proses klasifikasi teks karena sederhana, cepat, dan mampu menghasilkan performa yang baik meskipun menggunakan dataset yang relatif kecil [13].

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (3)$$

Keterangan:
P(C|X) = probabilitas data Xtermasuk ke dalam kelas C(*posterior probability*).
P(X|C) = probabilitas kemunculan data X pada kelas C(*likelihood*).
P(C) = probabilitas awal dari kelas C (*prior probability*).
P(X) = probabilitas total dari data X.
C = kelas atau kategori data.
X = data atau fitur yang akan diklasifikasikan.

E. Pengujian Sistem

$$Accuracy = \frac{Jumlah\ Data\ Benar}{Total\ Data} \times 100\% \quad (4)$$

Keterangan:
TP (True Positive) = data positif yang diprediksi benar.
TN (True Negative) = data negatif yang diprediksi benar.
FP (False Positive) = data negatif tetapi diprediksi positif.
FN (False Negative) = data positif tetapi diprediksi negative.

$$Precision = \frac{TP}{TP+FN} \quad (5)$$

Keterangan:
TP = True Positive
FP = False Positive

$$Recall = \frac{TP}{TP+FP} \quad (6)$$

Keterangan:
TP = True Positive
FN = False Negative

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Keterangan:
Precision = tingkat ketepatan prediksi
Recall = tingkat keberhasilan menemukan data benar

Pengujian sistem dilakukan untuk mengetahui performa model dalam melakukan klasifikasi otomatis jenis surat. Evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, dan *f1-score*. Nilai akurasi digunakan untuk mengetahui tingkat ketepatan model dalam mengklasifikasikan dokumen surat [14][15]. Hasil pengujian digunakan untuk mengetahui efektivitas metode TF-IDF dan algoritma Naive Bayes dalam proses klasifikasi otomatis jenis surat pada sistem e-arsip BPS Kabupaten Sragen.

III. HASIL DAN PEMBAHASAN

A. Hasil Pengujian Model

Pengujian model dilakukan menggunakan metode train test split dengan pembagian data sebesar 80% data latih dan 20% data uji. Proses klasifikasi dilakukan menggunakan metode TF-IDF dan algoritma Naive Bayes untuk menentukan jenis surat secara otomatis.

Hasil pengujian model menunjukkan bahwa sistem mampu melakukan klasifikasi surat dengan baik. Berdasarkan hasil pengujian diperoleh nilai akurasi sebesar 85,71%. Selain itu, diperoleh nilai *precision*, *recall*, dan *f1-score* untuk setiap kategori surat seperti ditunjukkan pada Tabel 2.

Tabel 1. Hasil Pengujian Model

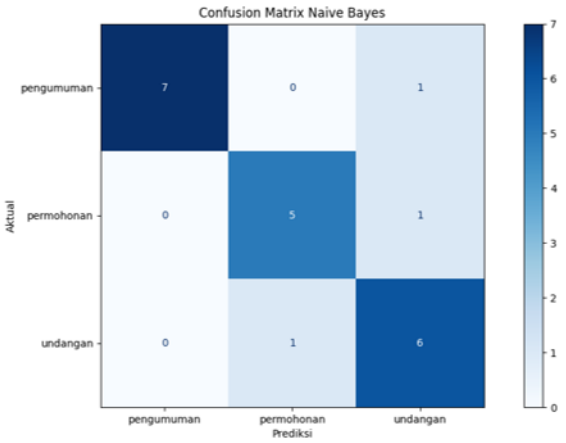
Kategori	Precision	Recall	F1-Score	Support
Pengumuman	1.00	0.88	0.93	8

Permohonan	0.83	0.83	0.83	6
Undangan	0.75	0.86	0.80	7
Accuracy			0.86	21

Berdasarkan hasil pengujian, kategori surat pengumuman memiliki nilai *precision* tertinggi sebesar 1.00 dengan nilai *f1-score* sebesar 0.93. Hal ini menunjukkan bahwa model mampu mengenali pola kata pada surat pengumuman dengan baik. Sementara itu, kategori surat undangan memiliki nilai *precision* paling rendah sebesar 0.75 karena masih terdapat beberapa kesalahan klasifikasi antar kategori surat.

B. Confusion Matrix

Hasil confusion matrix pada penelitian ini ditunjukkan pada Gambar 3.



Gambar 3. Confusion matrix

Hasil tersebut menunjukkan bahwa sebagian besar data berhasil diklasifikasikan sesuai kategori sebenarnya. Nilai pada diagonal *confusion matrix* menunjukkan jumlah prediksi yang benar, sedangkan nilai di luar diagonal menunjukkan kesalahan klasifikasi yang dilakukan model.

C. Perhitungan Manual TF-IDF

Perhitungan TF-IDF dilakukan untuk memberikan bobot pada setiap kata dalam dokumen surat. Sebagai contoh digunakan kalimat berikut, "Dengan hormat kami mengundang saudara untuk menghadiri rapat koordinasi"

Hasil *preprocessing* menghasilkan beberapa kata penting seperti yang terlihat pada Tabel 3.

Tabel 2. Kata Penting

Term	Tf
hormat	1
Undang	1
Hadir	1
Rapat	1
Koordinasi	1

Misalkan jumlah total dokumen adalah 105 dokumen dan kata "rapat" muncul pada 20 dokumen, maka nilai *IDF* dihitung menggunakan persamaan berikut:

Perhitungan kata "rapat":

$$IDF = log\frac{105}{20} \quad (8)$$

$$IDF = \log 5.25 \quad (9)$$

$$IDF = 0.72 \quad (10)$$

Karena nilai TF kata "rapat" 1, maka:

$$TF - IDF = 1 \times 0.72 \quad (11)$$

$$TF - IDF = 0.72 \quad (12)$$

Nilai tersebut menunjukkan tingkat kepentingan kata "rapat" pada dokumen surat.

D. Perhitungan Manual Naive Bayes

Perhitungan Naive Bayes dilakukan untuk menentukan probabilitas kategori surat berdasarkan kata yang muncul pada dokumen. Misalkan terdapat data sebagai berikut:

a. Total surat pengumuman = 39
b. Total surat undangan = 35
c. Total surat permohonan = 31
d. Total seluruh dokumen = 105

Probabilitas kategori pengumuman:

$$P(Pengumuman) = \frac{39}{105} = 0.37 \quad (13)$$

Probabilitas kategori Undangan:

$$P(Undangan) = \frac{35}{105} = 0.33 \quad (14)$$

Probabilitas kategori Permohonan:

$$P(Permohonan) = \frac{31}{105} = 0.30 \quad (15)$$

Misalkan terdapat dokumen uji dengan kata "mengundang rapat koordinasi"
Kemudian dihitung probabilitas setiap kata terhadap kategori surat. Contoh probabilitas pada kategori undangan seperti pada Tabel 4.

Tabel 3. Tabel probabilitas

Kata	Probabilitas
mengundang	0.50
rapat	0.40
koordinasi	0.35

Maka probabilitas kategori undangan dihitung sebagai berikut:

$$P(H|X) = P(H) \times P(x_1|H) \times P(x_2|H) \times P(x_3|H) \quad (16)$$

$$P(Undangan|X) = 0.33 \times 0.50 \times 0.40 \times 0.35 \quad (17)$$

$$P(Undangan)|X = 0.0231 \quad (18)$$

Hasil probabilitas terbesar akan menjadi hasil klasifikasi surat. Berdasarkan perhitungan tersebut, dokumen diprediksi termasuk kategori surat undangan karena memiliki probabilitas tertinggi dibanding kategori lainnya.

E. Perhitungan Manual Accuracy

Jumlah prediksi benar diperoleh dari penjumlahan diagonal *confusion matrix*:

a. Pengumuman = 7
b. Permohonan = 5
c. Undangan = 6

Total prediksi benar:

$$7 + 5 + 6 = 18 \quad (19)$$

Total data uji = 21

$$Accuracy = \frac{18}{21} \times 100\% = 85.7\% \quad (20)$$

Hasil perhitungan manual menunjukkan nilai akurasi sebesar 85,71%, sesuai dengan hasil pengujian sistem.

F. Perhitungan Manual Precision

Perhitungan *precision* kategori pengumuman:

$$Precision = \frac{7}{7+0} = 1.00 \quad (21)$$

Perhitungan *precision* kategori permohonan:

$$Precision = \frac{5}{5+1} = 0.83 \quad (22)$$

Perhitungan *precision* kategori undangan:

$$Precision = \frac{6}{6+2} = 0.75 \quad (23)$$

G. Perhitungan Manual Recall

Perhitungan *recall* kategori pengumuman:

$$Recall = \frac{7}{7+1} = 0.88 \quad (24)$$

Perhitungan *recall* kategori permohonan:

$$Recall = \frac{5}{5+1} = 0.83 \quad (25)$$

Perhitungan *recall* kategori undangan:

$$Recall = \frac{6}{6+1} = 0.86 \quad (26)$$

H. Perhitungan Manual F1-Score

Perhitungan *F1-score* kategori pengumuman:

$$F1 = 2 \times \frac{1.00 \times 0.88}{1.00+0.88} = 0.93 \quad (27)$$

Perhitungan *F1-score* kategori permohonan:

$$F1 = 2 \times \frac{0.83 \times 0.83}{0.83+0.83} = 0.83 \quad (28)$$

Perhitungan F1-*score* kategori undangan:

$$F1 = 2 \times \frac{0.75 \times 0.86}{0.75+0.86} = 0.80 \quad (29)$$

I. Implementasi Antarmuka Sistem

Implementasi antarmuka sistem dilakukan pada aplikasi e-arsip berbasis web yang dikembangkan menggunakan *framework* Flask untuk mendukung pengelolaan arsip surat dan klasifikasi otomatis jenis surat.

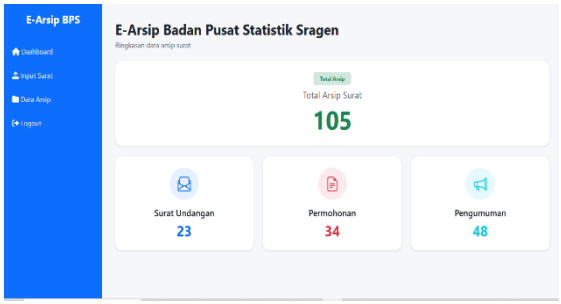
a. Halaman Login

Halaman login digunakan untuk autentikasi pengguna sebelum mengakses sistem. Tampilan halaman login dapat dilihat pada Gambar 4.

Gambar 4. Halaman login

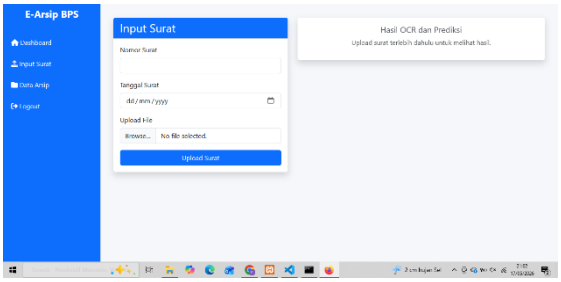
b. Halaman Dashboard

Halaman dashboard menampilkan menu utama dan informasi pengelolaan arsip surat. Tampilan halaman dashboard dapat dilihat pada Gambar 5.



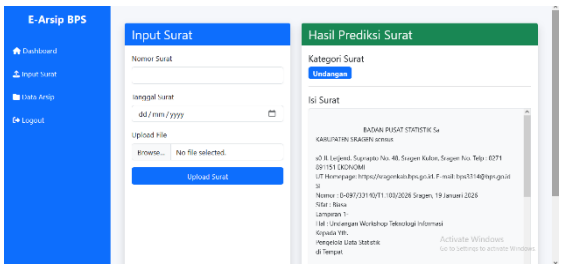
Gambar 5. Halaman dashboard

c. Halaman Upload Surat
Halaman upload surat digunakan untuk mengunggah dokumen yang akan diproses oleh sistem. Tampilan halaman upload surat dapat dilihat pada Gambar 6.



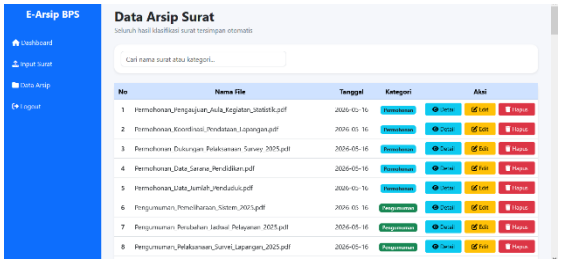
Gambar 6. Halaman upload surat

d. Halaman Hasil Klasifikasi
Halaman hasil klasifikasi menampilkan kategori surat berdasarkan hasil prediksi sistem. Tampilan halaman hasil klasifikasi dapat dilihat pada Gambar 7.



Gambar 7. Halaman hasil klasifikasi

e. Halaman Arsip Surat
Halaman arsip surat digunakan untuk menampilkan data surat yang tersimpan pada sistem. Tampilan halaman arsip surat dapat dilihat pada Gambar 8.



Gambar 8. Halaman arsip surat

IV. KESIMPULAN

Document Frequency (TF-IDF) dan algoritma Multinomial Naive Bayes berhasil diterapkan pada sistem e-arsip BPS Kabupaten Sragen untuk melakukan klasifikasi otomatis jenis surat. Hasil pengujian menunjukkan bahwa model mampu mengklasifikasikan surat ke dalam kategori pengumuman, undangan, dan permohonan dengan nilai akurasi sebesar 85,71%. Selain itu, hasil evaluasi menggunakan metrik *precision*, *recall*, dan *F1-score* menunjukkan performa klasifikasi yang baik, dimana kategori pengumuman memperoleh nilai *precision* 1,00, *recall* 0,88, dan *F1-score* 0,93, sedangkan kategori permohonan dan undangan juga menunjukkan hasil yang cukup baik. Hasil tersebut menunjukkan bahwa kombinasi metode TF-IDF dan algoritma Multinomial Naive Bayes mampu membantu proses pengelolaan arsip surat secara lebih cepat, efektif, dan efisien. Implementasi sistem berbasis web menggunakan framework Flask juga mempermudah pengguna dalam melakukan unggah, pencarian, dan penyimpanan arsip surat secara digital.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Badan Pusat Statistik (BPS) Kabupaten Sragen serta semua pihak yang telah memberikan dukungan dan bantuan dalam pelaksanaan penelitian ini.

REFERENSI

[1] Mujaki, A.F., 2026. Klasifikasi Jenis Surat pada Sistem E-Arsip Menggunakan Metode K-Nearest Neighbor. Jurnal Ilmiah Sistem Informasi, 5(1). https://doi.org/10.51903/336mw957.
[2] Junaidi, I.H., Sopingi and Sumarlinda, S., 2024. Sistem Retrieval E-Arsip Tirta Asata Menggunakan Algoritma Vector Space Model.

Jurnal Fasilkom, 14(2), pp.361–369. https://doi.org/10.37859/jf.v14i2.7358.

[3] Gerung, F.R., Maramis, G.D.P. and Moningkey, E.R.S., 2025. Penerapan Algoritma Naive Bayes pada Arsip Surat Kantor Kecamatan Motoling Barat. REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer, 9(2), pp.676–690. https://doi.org/10.33395/remik.v9i2.14786.

[4] Umar, N. and Nur, M.A., 2022. Application of Naïve Bayes Algorithm Variations on Indonesian General Analysis Dataset for Sentiment Analysis. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6(4), pp.585–590. https://doi.org/10.29207/resti.v6i4.4179.

[5] Manikome, H. and Maramis, G.D.P., 2025. Arsip Surat Masuk dan Keluar pada Kejaksaan Tinggi Sulawesi Utara dengan Algoritma K-Means Berbasis Web. Journal of Informatics, Business, Education and Innovation Technology, 3(3). https://jibeit.teknikinformatika.org/index.php/jibeit/article/view/218.

[6] Pattikawa, S.J. and Sundari, O., 2024. Penerapan Manajemen Arsip Elektronik dalam Optimalisasi Pengelolaan Arsip di PT Kreativitas Ganesha Sejahtera. Jurnal Economic, Business and Accounting, 7(5), pp.3682–3702. https://doi.org/10.31539/costing.v7i5.11862.

[7] Kurniawan, D.L., Immasari, I.R. and Sianipar, A.Z., 2022. Perancangan Sistem Informasi Pengarsipan Berbasis Website. Jurnal Manajemen Informatika Jayakarta, 2(1), p.77. https://doi.org/10.52362/jmijayakarta.v2i1.685.

[8] Faturrahman, F., 2025. Implementasi Text Mining dengan TF-IDF untuk Klasifikasi Jenis Perbaikan Kendaraan dan Evaluasi Akurasi Algoritma Machine Learning. Undergraduate thesis. Universitas Islam Indonesia. https://dspace.uii.ac.id/123456789/59697.

[9] Gadewar, S.A. and Pawar, P.H., 2024. Natural Language Processing and Deep Learning Approaches for Multiclass Document Classifier. International Journal of Scientific Research in Science, Engineering and Technology, pp.278–283. https://doi.org/10.32628/IJSRSET2411143.

[10] Aulia, D.W.A.P., 2025. Text Mining dari Jurnal Teknologi Pendidikan Menggunakan Term Frequency Matrix. Science Journal of Statistics and Probability Its Applications, 3(1), pp.15–28. https://doi.org/10.24127/sciencestatistics.v3i1.5989.

[11] Mi'raj Fattah, N.A.V. and Siswa, T.A.Y., 2026. Analisis Komparasi TF-IDF dan Word2Vec pada Algoritma SVM untuk Sentimen Pembangunan IKN di YouTube. *BUFFER Informatika*, 12(1). https://doi.org/10.25134/buffer.v12i1.551.

[12] Comparison and Analysis of the Accuracy of Multiple Machine Learning Algorithms in the Field of Spam Classification, 2023. https://doi.org/10.1117/12.2675259.

[13] Ionendri, N. A., Candra, F., & Rizal, A., "News Classification using Natural Language Processing with TF-IDF and Multinomial Naïve Bayes," Journal of Applied Computer Science and Technology, vol. 6, no. 1, pp. 37–45, 2025. https://doi.org/10.52158/jacost.v6i1.1099

[14] Indey, J.F. and Supangat, S., 2024. Implementasi Algoritma Naïve Bayes dalam Sistem Pengarsipan Surat Berbasis AI di GPI Papua Klasis Mimika Papua Tengah. *Jurnal Ilmiah Informatika,* 12(2), pp.102–113. https://doi.org/10.33884/jif.v12i02.9087.

[15] Yuyun, Y., 2023. Klasifikasi Surat Digital Menggunakan Algoritma *Machine Learning*. *Jurnal IT*, 13(2), pp.66–71. https://doi.org/10.37639/jti.v13i2.350.