KLASIFIKASI SURAT DINAS BERBASIS TEXT MINING MENGGUNAKAN TF-IDF DAN SUPPORT VECTOR MACHINE

Falih Walen Satria Shakti, Eko Purwanto, Afu Ichsan Pradana
Teknik Informatika, Universitas Duta Bangsa Surakarta
Jalan Bhayangkara No.55, Tipes, Kec. Serengan, Kota Surakarta, Jawa Tengah
afu_ichsan@udb.ac.id

ABSTRAK

Pengelolaan surat dinas pada instansi pemerintahan memerlukan proses klasifikasi yang tepat untuk mendukung efektivitas administrasi dan pengarsipan dokumen. Namun, proses klasifikasi surat yang masih dilakukan secara manual sering menimbulkan ketidakkonsistenan, keterlambatan pencarian arsip, dan tingginya potensi kesalahan pengelompokan dokumen. Penelitian ini bertujuan menerapkan metode TF-IDF dan Support Vector Machine (SVM) untuk mengklasifikasikan surat dinas berdasarkan teks pada bagian perihal surat. Dataset penelitian berasal dari 2.155 dokumen surat dinas Sekretariat Dinas Komunikasi dan Informatika Kabupaten Sragen yang dikumpulkan dari periode Januari hingga Desember 2025. Setelah melalui proses preprocessing dan seleksi data, diperoleh 188 dokumen valid yang dikelompokkan ke dalam 15 kategori surat. Tahapan penelitian meliputi preprocessing teks berupa case folding, tokenizing, stopword removal, dan stemming, kemudian dilanjutkan dengan pembobotan TF-IDF berbasis n-gram (1,2) serta klasifikasi menggunakan algoritma SVM kernel Radial Basis Function (RBF). Evaluasi model dilakukan menggunakan pembagian data 80:20 dan 10-fold cross-validation dengan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model memperoleh akurasi sebesar 76,32% pada skenario 80:20 dan rata-rata 72,98% pada 10-fold cross-validation. Hasil tersebut menunjukkan bahwa metode TF-IDF dan SVM mampu digunakan untuk klasifikasi surat dinas secara cukup baik.

Kata kunci : *Text Mining, TF-IDF, Support Vector Machine, Klasifikasi Teks, Surat Dinas*

1. PENDAHULUAN

Pengelolaan surat dinas memiliki peran penting dalam mendukung kelancaran proses administrasi pada instansi pemerintahan. Surat dinas digunakan sebagai media komunikasi resmi dalam penyampaian informasi, pemberian instruksi, pengambilan keputusan, penugasan, hingga pengarsipan dokumen. Pengelolaan surat yang dilakukan secara terstruktur dapat membantu instansi dalam mempercepat proses pencarian dokumen, mempermudah disposisi surat, serta menjaga keteraturan arsip. Sebaliknya, pengelolaan surat yang tidak terorganisasi dapat menyebabkan keterlambatan administrasi, kesalahan penyimpanan dokumen, hingga menurunnya efektivitas pelayanan administrasi pemerintahan. Pemanfaatan teknologi kecerdasan buatan dalam pengelolaan arsip pemerintahan juga mulai banyak dikembangkan untuk meningkatkan aksesibilitas dan efisiensi pengelolaan dokumen digital[1].

Namun, proses pengelompokan surat dinas pada banyak instansi pemerintahan masih dilakukan secara manual oleh pegawai administrasi. Proses tersebut dilakukan dengan membaca isi atau perihal surat kemudian menentukan kategori surat berdasarkan pemahaman petugas. Cara tersebut dinilai kurang efektif karena sangat bergantung pada ketelitian dan konsistensi petugas administrasi. Selain itu, meningkatnya jumlah surat masuk dan surat keluar setiap periode menyebabkan proses pengelompokan dokumen memerlukan waktu yang lebih lama dan berpotensi menimbulkan kesalahan klasifikasi. Kondisi tersebut juga terjadi pada Sekretariat Dinas Komunikasi dan Informatika Kabupaten Sragen, dimana proses klasifikasi surat masih dilakukan secara manual sehingga pengelolaan dokumen belum berjalan secara optimal. Oleh karena itu, diperlukan suatu sistem yang mampu melakukan klasifikasi surat secara otomatis agar proses pengelolaan arsip dapat dilakukan secara lebih cepat, konsisten, dan efisien.

Perkembangan teknologi di bidang text mining dan machine learning memberikan peluang dalam menyelesaikan permasalahan klasifikasi dokumen teks secara otomatis. Text mining merupakan metode yang digunakan untuk mengekstraksi informasi penting dari data teks tidak terstruktur sehingga dapat diolah menjadi informasi yang lebih terorganisasi. Dalam proses klasifikasi teks, tahapan preprocessing memiliki peran penting untuk meningkatkan kualitas data sebelum dilakukan pemodelan. Tahapan preprocessing meliputi case folding untuk menyeragamkan huruf menjadi huruf kecil, tokenizing untuk memecah kalimat menjadi token kata, stopword removal untuk menghapus kata yang tidak memiliki makna penting, serta stemming untuk mengubah kata menjadi bentuk dasar. Tahapan tersebut bertujuan untuk mengurangi noise pada data teks sehingga informasi yang digunakan dalam proses klasifikasi menjadi lebih relevan[2].

Setelah proses preprocessing dilakukan, dokumen teks perlu direpresentasikan ke dalam bentuk numerik agar dapat diproses oleh algoritma machine learning. Salah satu metode representasi fitur yang banyak digunakan adalah Term Frequency–Inverse Document Frequency (TF-IDF). Metode TF-IDF

bekerja dengan memberikan bobot pada setiap kata berdasarkan tingkat kemunculan kata dalam dokumen dan tingkat kepentingannya terhadap keseluruhan dokumen. Kata yang sering muncul pada suatu dokumen namun jarang ditemukan pada dokumen lain akan memiliki bobot lebih tinggi sehingga dianggap lebih representatif dalam proses klasifikasi. Selain itu, penggunaan fitur berbasis n-gram dapat membantu sistem dalam mengenali pola hubungan antar kata pada teks perihal surat sehingga informasi kontekstual pada dokumen dapat direpresentasikan dengan lebih baik. Penelitian sebelumnya menunjukkan bahwa penggunaan TF-IDF mampu meningkatkan performa klasifikasi teks dibandingkan metode representasi sederhana seperti Bag-of-Words[3].

Salah satu algoritma machine learning yang banyak digunakan dalam klasifikasi teks adalah Support Vector Machine (SVM). Algoritma ini memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan sparse seperti data teks. SVM bekerja dengan membentuk hyperplane optimal yang mampu memisahkan data ke dalam kelas yang berbeda dengan margin maksimum sehingga menghasilkan performa klasifikasi yang baik [4].

Selain itu, penggunaan kernel pada SVM memungkinkan model untuk menangani data yang tidak dapat dipisahkan secara linear. Kernel Radial Basis Function (RBF) merupakan salah satu kernel yang sering digunakan karena mampu memetakan data ke dimensi yang lebih tinggi sehingga pola data yang kompleks dapat dipisahkan dengan lebih optimal. Beberapa penelitian menunjukkan bahwa kombinasi TF-IDF dan SVM mampu memberikan performa klasifikasi yang baik pada berbagai kasus klasifikasi teks[5].

Berbagai penelitian mengenai klasifikasi teks telah banyak dilakukan menggunakan kombinasi metode TF-IDF dan algoritma machine learning, seperti klasifikasi berita, analisis media sosial, maupun ulasan produk. Namun, sebagian besar penelitian masih berfokus pada dokumen teks umum dan belum banyak diterapkan pada dokumen administratif pemerintahan, khususnya surat dinas berdasarkan teks pada bagian perihal surat. Padahal, surat dinas memiliki karakteristik bahasa yang lebih formal dan penggunaan istilah administratif yang berbeda dibandingkan dokumen teks umum lainnya. Oleh karena itu, diperlukan penelitian yang secara khusus membahas penerapan text mining untuk klasifikasi surat dinas agar dapat mendukung pengelolaan administrasi pemerintahan secara lebih efektif dan terstruktur[1].

Berdasarkan permasalahan tersebut, penelitian ini menerapkan pendekatan text mining menggunakan tahapan preprocessing teks, pembobotan fitur menggunakan metode TF-IDF berbasis n-gram, serta klasifikasi menggunakan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF). Proses evaluasi model dilakukan menggunakan metode pembagian data 80:20 dan 10-fold cross-validation untuk mengukur performa dan kestabilan model klasifikasi. Evaluasi performa model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score untuk mengetahui kemampuan model dalam mengklasifikasikan surat dinas secara akurat dan konsisten.

Tujuan dari penelitian ini adalah menerapkan metode TF-IDF dan Support Vector Machine (SVM) untuk mengklasifikasikan surat dinas berdasarkan teks pada bagian perihal surat di Sekretariat Dinas Komunikasi dan Informatika Kabupaten Sragen. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi performa model klasifikasi menggunakan pengujian accuracy, precision, recall, dan F1-score. Hasil penelitian diharapkan dapat membantu meningkatkan efisiensi pengelolaan arsip surat dinas, mempercepat proses klasifikasi dokumen, serta memberikan kontribusi dalam pengembangan penerapan text mining pada dokumen administratif pemerintahan.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Berbagai penelitian terdahulu telah mengkaji penerapan metode klasifikasi teks menggunakan pendekatan text mining, pembobotan TF-IDF, serta algoritma Support Vector Machine (SVM)[5]. Penelitian-penelitian tersebut menjadi landasan sekaligus pembanding untuk mengukur kontribusi penelitian ini secara ilmiah. Tabel 1 menyajikan ringkasan penelitian terdahulu yang relevan dengan topik penelitian ini.

Tabel 1. Perbandingan penelitian terdahulu

No	Peneliti & Tahun	Judul	Metode	Hasil Akurasi	Perbedaan dengan Penelitian Ini
[1]	Das et al. (2023)	A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset	TF-IDF + SVM, Random Forest, Naïve Bayes	93,81% (Random Forest)	Menggunakan dataset ulasan produk; penelitian ini fokus pada surat dinas pemerintahan berbahasa Indonesia
[2]	Al-Habib et al. (2023)	Text Processing Using Support Vector Machine for Scientific Research Paper Classification	TF-IDF + SVM	Akurasi tinggi (SVM optimal)	Fokus pada dokumen ilmiah; penelitian ini menggunakan dokumen administratif surat dinas
[3]	Tsai & Chang (2023)	Efficient Selection of Gaussian Kernel SVM Parameters for Imbalanced	SVM-RBF kernel + parameter	Bervariasi per dataset	Berfokus pada dataset biomedis; penelitian ini menguji kernel RBF pada teks

No	Peneliti & Tahun	Judul	Metode	Hasil Akurasi	Perbedaan dengan Penelitian Ini
		Data	tuning		surat dinas yang tidak seimbang
[4]	Aubaid et al. (2024)	Text Classification using Machine Learning and Embedding Techniques	TF-IDF + Machine Learning + Embedding	Akurasi tinggi	Menggunakan berbagai metode; penelitian ini fokus pada TF-IDF dan SVM
[5]	Piasta & Kotas (2025)	Comparative Analysis of Natural Language Processing Techniques in the Classification of Press Articles	TF-IDF, N-Gram, Machine Learning	OpenNLP 84%, Weka 86%, CoreNLP 88%, DistilBERT 92%.	Penelitian menggunakan artikel pers, sedangkan penelitian ini menggunakan surat dinas pemerintahan
[6]	Karim et al. (2024)	Strengthening Fake News Detection using SVM and TF-IDF	TF-IDF + SVM	Akurasi tinggi	Fokus pada deteksi berita;Penelitian ini fokeus pada klasifikasi surat dinas.

Berdasarkan Tabel 1, dapat disimpulkan bahwa penelitian ini memiliki keunikan tersendiri dibandingkan penelitian terdahulu. Pertama, objek penelitian ini adalah dokumen surat dinas pemerintahan berbahasa Indonesia yang dikumpulkan dari instansi nyata (Sekretariat Diskominfo Kabupaten Sragen), sehingga penelitian ini memiliki nilai aplikatif yang tinggi [1].

Kedua, penelitian ini mengklasifikasikan 15 kategori surat yang mencerminkan kompleksitas administrasi pemerintahan. Ketiga, penggunaan n-gram (1,2) pada representasi TF-IDF memberikan kemampuan menangkap frasa kontekstual yang khas dalam dokumen administratif [6][7].

Keempat, evaluasi dilakukan secara komprehensif menggunakan dua metode, yaitu pemisahan 80:20 dan 10-fold cross validation, sehingga hasil yang diperoleh lebih andal secara statistik.

2.2. Text Mining

Text mining adalah suatu metode yang digunakan untuk mengolah dan menganalisis data berbentuk teks yang tidak terstruktur dengan tujuan menemukan pola, keterkaitan, serta informasi penting di dalamnya. Proses ini dilakukan dengan mengubah teks menjadi bentuk data yang lebih terstruktur agar dapat dianalisis secara sistematis.[2].

Secara umum pada tahapan dalam text mining meliputi pengumpulan data, preprocessing, ekstraksi fitur, pemodelan, hingga evaluasi. Pada penelitian ini, text mining diterapkan pada teks bagian perihal surat dinas untuk mengklasifikasikan dokumen secara otomatis. Pendekatan ini dipilih karena pengolahan secara manual sering kali tidak konsisten dan memerlukan waktu yang lebih lama, sehingga diperlukan metode berbasis machine learning untuk meningkatkan akurasi serta efisiensi dalam pengolahan dokumen[8].

2.3. Preprocessing Teks

Preprocessing merupakan tahap awal dalam text mining yang bertujuan untuk menyiapkan data teks agar lebih bersih dan terstruktur sebelum dilakukan analisis lebih lanjut. Tahap ini memiliki peran penting karena secara langsung memengaruhi kualitas hasil klasifikasi yang dihasilkan oleh model.Tahapan preprocessing yang diterapkan meliputi:

a. Case Folding, yaitu proses mengubah seluruh karakter teks menjadi huruf kecil (lowercase).
b. Tokenizing, yaitu proses memisahkan string teks menjadi unit-unit kata yang disebut token.
c. Stopword Removal, yaitu penghapusan kata-kata umum yang tidak memiliki nilai informatif signifikan bagi proses klasifikasi.
d. Stemming, yaitu proses mereduksi kata berimbuhan menjadi kata dasar (root word).

2.4. Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF merupakan salah satu teknik pembobotan kata yang umum digunakan dalam text mining dan information retrieval. Metode ini berfungsi untuk menentukan tingkat kepentingan suatu kata dalam sebuah dokumen dengan membandingkannya terhadap seluruh kumpulan dokumen (corpus), sehingga kata yang lebih relevan akan memiliki bobot yang lebih tinggi. TF-IDF menggabungkan dua komponen utama, yaitu:

a. Term Frequency (TF), yang mengukur frekuensi kemunculan suatu kata dalam sebuah dokumen.
b. Inverse Document Frequency (IDF), yang mengukur keunikan atau kepentingan umum suatu kata dalam seluruh corpus.

Nilai TF-IDF dihitung menggunakan persamaan: $TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$, di mana $IDF(t) = \log(N/df(t))$, dengan N adalah jumlah total dokumen dan df(t) adalah jumlah dokumen yang mengandung kata t [3]. Penelitian sebelumnya menunjukkan bahwa TF-IDF secara signifikan meningkatkan performa klasifikasi dibandingkan pendekatan Bag-of-Words (BoW) sederhana karena kemampuannya dalam mengidentifikasi kata-kata yang paling diskriminatif untuk setiap kelas.

Penggunaan n-gram bersama TF-IDF memberikan kemampuan tambahan dalam menangkap frasa multi-kata yang bermakna[9].

N-gram (1,2) atau bigram memungkinkan model untuk mempertimbangkan frasa dua kata sebagai satu

unit fitur, sehingga konteks semantik antar kata dapat direpresentasikan dengan lebih baik[10] .

2.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang bekerja dengan mencari hyperplane optimal untuk memisahkan data ke dalam kelas yang berbeda dengan margin maksimum [4]. SVM dikenal efektif dalam menangani data berdimensi tinggi seperti data teks. Pada data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk memetakan data ke ruang berdimensi lebih tinggi. Dalam penelitian ini digunakan kernel Radial Basis Function (RBF) karena mampu menangani data yang kompleks dan tidak linear. Untuk klasifikasi multi-kelas, digunakan pendekatan One-vs-Rest[11].

2.6. Evaluasi Model Klasifikasi

Evaluasi model dilakukan untuk mengukur kinerja algoritma dalam melakukan klasifikasi. Metode evaluasi yang digunakan adalah confusion matrix yang menghasilkan beberapa metrik, yaitu:

a. Accuracy untuk mengukur tingkat ketepatan prediksi secara keseluruhan
b. Precision untuk mengukur ketepatan prediksi pada kelas tertentu
c. Recall untuk mengukur kemampuan model dalam menemukan data yang relevan
d. F1-score sebagai kombinasi antara precision dan recall

Penggunaan beberapa metrik ini penting untuk memberikan gambaran performa model secara menyeluruh, terutama pada data yang tidak seimbang.

2.7. K-Fold Cross Validation

K-fold cross validation merupakan metode evaluasi yang dilakukan dengan membagi dataset menjadi beberapa bagian (fold), di mana setiap bagian digunakan secara bergantian sebagai data uji dan sisanya sebagai data latih[12].

Metode ini bertujuan untuk menghasilkan evaluasi yang lebih stabil serta mengurangi bias akibat pembagian data tertentu[13].

Dalam penelitian ini digunakan 10-fold cross validation untuk memperoleh hasil evaluasi model yang lebih akurat dan representatif.

2.8. Manajemen Arsip dan Surat Dinas

Surat dinas merupakan dokumen resmi dalam administrasi pemerintahan yang perlu dikelola secara sistematis sebagai bagian dari manajemen arsip untuk menjamin keteraturan dan kemudahan akses informasi. Namun, pengelolaan yang masih dilakukan secara manual sering menimbulkan permasalahan seperti ketidakkonsistenan klasifikasi, keterlambatan pengolahan, serta kesulitan dalam pencarian arsip. Oleh karena itu, diperlukan sistem klasifikasi otomatis berbasis text mining dan machine learning untuk meningkatkan efisiensi, konsistensi, dan akurasi dalam pengelolaan surat dinas secara optimal[14].

3. METODE PENELITIAN

Pada bagian ini memuat metode yang digunakan pada pembuatan skripsi.

3.1. Dataset Penelitian

Dataset dalam penelitian ini berupa kumpulan dokumen surat dinas yang berasal dari arsip Sekretariat Dinas Komunikasi dan Informatika Kabupaten Sragen. Data diperoleh secara manual dari dokumen surat yang diterbitkan selama periode Januari hingga Desember 2025. Secara keseluruhan, dataset awal terdiri dari 2.155 dokumen surat dinas.

Distribusi data awal berdasarkan bulan penerbitan disajikan pada Tabel 2. Distribusi ini menunjukkan bahwa jumlah surat pada setiap bulan tidak seragam. Jumlah dokumen terbanyak terdapat pada bulan Maret sebanyak 283 dokumen (13,1%) dan Juli sebanyak 279 dokumen (12,9%), sedangkan jumlah dokumen paling sedikit terdapat pada bulan April sebanyak 73 dokumen (3,4%) dan Desember sebanyak 77 dokumen (3,6%).

Tabel 2. Distribusi dataset

Bulan	Jumlah	persentase
Januari	194	9,0%
Februari	199	9,2%
Maret	283	13,1%
April	73	3,4%
Mei	256	11,9%
Juni	179	8,3%
Juli	279	12,9%
Agustus	101	4,7%
September	184	8,5%
Oktober	152	7,1%
November	178	8,3%
Desember	77	3,6%
Total	2.155	100%

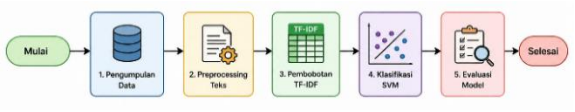
Setelah dilakukan proses preprocessing dan seleksi data berdasarkan kelengkapan isi perihal surat serta kesesuaian kategori, diperoleh 188 dokumen valid yang digunakan dalam proses klasifikasi. Data tersebut kemudian dikelompokkan ke dalam 15 kategori surat, seperti SPPD, SOP, Surat Tugas, Permohonan, Undangan, dan kategori administratif lainnya.

Dataset selanjutnya dibagi menggunakan rasio 80:20 dengan teknik stratified split untuk menjaga proporsi distribusi kelas pada data latih dan data uji. Sebanyak 150 dokumen digunakan sebagai data latih dan 38 dokumen digunakan sebagai data uji. Setelah melalui proses preprocessing, setiap dokumen memiliki rata-rata 3–4 token dengan panjang minimum 1 token dan maksimum 25 token.

3.2. Tahapan Penelitian

Berdasarkan dataset 2.155 dokumen surat dinas yang telah diuraikan pada subbab 3.1, penelitian ini

dilaksanakan melalui lima tahap yang berurutan dan saling bergantung, sebagaimana ditunjukkan pada Gambar 1. Setiap tahap dirancang untuk memastikan data teks yang diolah memenuhi standar kualitas sebelum dimasukkan ke dalam model klasifikasi.



Gambar 1. Alur tahapan penelitian

3.3. Metode yang Digunakan

Metode yang digunakan dalam penelitian ini adalah pembobotan fitur menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dan klasifikasi menggunakan Support Vector Machine (SVM). Dataset penelitian berasal dari 2.155 dokumen surat dinas Sekretariat Dinas Komunikasi dan Informatika Kabupaten Sragen yang dikumpulkan selama periode Januari hingga Desember 2025. Setelah melalui tahapan preprocessing dan seleksi data, diperoleh 188 dokumen valid yang digunakan dalam proses klasifikasi ke dalam 15 kategori surat berdasarkan teks pada bagian perihal surat. Tahapan preprocessing meliputi case folding, tokenizing, stopword removal, dan stemming, kemudian dilanjutkan dengan pembobotan TF-IDF berbasis n-gram (1,2) dan klasifikasi menggunakan SVM kernel Radial Basis Function (RBF). Evaluasi model dilakukan menggunakan pembagian data 80:20 dan 10-fold cross-validation untuk mengukur performa dan kestabilan model.

3.4. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data teks perihal surat dinas yang diperoleh dari Sekretariat Dinas Komunikasi dan Informatika (Diskominfo) Kabupaten Sragen. Pengumpulan data dilakukan melalui wawancara dengan pegawai sekretariat untuk memahami alur pengelolaan surat serta menentukan kategori surat yang akan digunakan sebagai label kelas. Data yang digunakan telah melalui proses anonimisasi untuk menjaga kerahasiaan informasi instansi.

3.5. Preprocessing Data

Data teks perihal surat dinas terlebih dahulu melalui tahap preprocessing sebelum diproses lebih lanjut. Tahapan preprocessing yang dilakukan meliputi:

a. Case folding, yaitu mengubah seluruh teks menjadi huruf kecil
b. Tokenizing, yaitu memisahkan teks menjadi satuan token kata
c. Stopword removal, yaitu menghapus kata-kata yang tidak memiliki makna signifikan
d. Stemming, yaitu mengubah kata berimbuhan menjadi kata dasar menggunakan algoritma Sastrawi.

3.6. Ekstraksi Fitur dengan Term Frequency-Inverse Document Frequency (TF-IDF)

Setelah preprocessing, data teks diubah menjadi representasi numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Nilai TF mengukur frekuensi kemunculan suatu kata dalam dokumen, sedangkan nilai IDF mengukur seberapa jarang kata tersebut muncul pada seluruh koleksi dokumen. Bobot TF-IDF dihitung menggunakan persamaan sebagai berikut.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$
$$IDF(t) = \log(N/df(t))$$

Keterangan :

t	= term (kata)
d	= dokumen
TF(t,d)	= frekuensi kemunculan term t dalam dokumen d
IDF(t)	= inverse document frequency dari term t,
N	= jumlah total dokumen
df(t)	= jumlah dokumen yang mengandung term t

3.7. Klasifikasi dengan Support Vector Machine

Proses klasifikasi surat dinas dilakukan menggunakan algoritma Support Vector Machine (SVM). Metode ini dipilih karena mampu menangani data berdimensi tinggi seperti teks serta menghasilkan pemisahan kelas yang optimal [15]. SVM bekerja dengan menentukan batas pemisah terbaik (hyperplane) yang dapat membedakan data ke dalam beberapa kategori dengan jarak pemisah yang maksimal [16]. Secara matematis, hyperplane dapat dituliskan dalam bentuk persamaan sederhana sebagai berikut:

$$\omega \cdot x + b = 0$$

Dimana ω merupakan bobot, x adalah data input, dan b adalah bias. Model SVM kemudian mencari nilai parameter tersebut agar mampu memisahkan data secara optimal antar kelas. Untuk menangani data yang tidak dapat dipisahkan secara linear, penelitian ini menggunakan kernel Radial Basis Function (RBF) yang memetakan data ke ruang berdimensi lebih tinggi, sehingga pola yang kompleks dapat dikenali dengan lebih baik.

3.3.1. Evaluasi Model

Evaluasi performa model klasifikasi dilakukan menggunakan metode k-fold cross validation. Metrik evaluasi yang digunakan meliputi accuracy, precision, recall, dan F1-score yang dihitung berdasarkan confusion matrix. Persamaan masing-masing metrik ditunjukkan pada persamaan (3) hingga (6).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$
$$Precision = TP/\ (TP + FP)$$
$$Recall = TP\ /\ (TP + FN)$$
$$F1 - Score = 2 \times (Precision \times Recall)/\ (Precision + Recall)$$

Keterangan :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berupa arsip surat keluar tahun 2025 yang diperoleh dari Sekretariat Dinas Komunikasi dan Informatika (Diskominfo) Kabupaten Sragen. Data dikumpulkan melalui metode wawancara dan dokumentasi langsung di instansi terkait. Dataset awal berjumlah 2.155 dokumen surat dinas, kemudian dilakukan proses pembersihan dan penyaringan berdasarkan kelengkapan data serta kesesuaian kategori. Hasilnya diperoleh 188 data valid yang digunakan dalam penelitian dan terbagi ke dalam 15 kategori surat. Distribusi kategori surat secara lengkap disajikan pada Tabel 3.

Tabel 3. Distribusi dataset per kategori surat

No	Kategori Surat	Jumlah	%
1	SPPD	50	26,60
2	SK (Surat Keputusan)	27	14,36
3	Surat Tugas	25	13,30
4	SOP	19	10,11
5	Berita Acara	10	5,32
6	Surat Pengantar	10	5,32
7	Permohonan	9	4,79
8	Surat Kepegawaian	8	4,26
9	Pengadaan	5	2,66
10	Rekomendasi	5	2,66
11	Undangan	5	2,66
12	Data Personel	4	2,13
13	Surat Keterangan	4	2,13
14	Surat Balasan	4	2,13
15	Laporan	3	1,60
	Total	**188**	**100**

Berdasarkan Tabel 3, distribusi data menunjukkan ketidakseimbangan kelas (class imbalance) yang cukup signifikan. Kategori SPPD mendominasi dengan 50 data (26,60%), sementara kategori Laporan hanya diwakili oleh 3 data (1,60%). Ketidakseimbangan ini merupakan cerminan nyata dari kondisi administrasi persuratan di instansi pemerintahan, di mana surat perjalanan dinas memiliki frekuensi paling tinggi. Kondisi ini menjadi tantangan tersendiri bagi model klasifikasi, namun sekaligus mempertegas relevansi dan keaslian dataset yang digunakan dalam penelitian ini.

4.2. Hasil Preprocessing

Seluruh data teks perihal surat melewati empat tahap preprocessing secara berurutan. Proses dimulai dari case folding untuk menyeragamkan huruf, dilanjutkan tokenizing, stopword removal menggunakan kamus PySastrawi, dan diakhiri stemming menggunakan algoritma Sastrawi. Tabel 2 menampilkan perbandingan teks sebelum dan sesudah preprocessing.

Tabel 4. Distribusi dataset per kategori surat

Teks Asli (Perihal Surat)	Hasil Preprocessing
SK perjanjian pelaksanaan pekerjaan	janji laksana kerja
SOP pemusnahan Arsip	sop musnah arsip
Surat tugas Monitoring jaringan TI	surat tugas monitoring jaring
BA rekonsiliasi BMD per 31 Desember	ba rekonsiliasi bmd
SPPD	sppd

Hasil preprocessing pada Tabel 4 menunjukkan bahwa proses stemming menggunakan algoritma Sastrawi bekerja dengan efektif dalam mereduksi variasi morfologis kata bahasa Indonesia. Frasa seperti "pelaksanaan pekerjaan" berhasil disederhanakan menjadi "laksana kerja", dan "pemusnahan" menjadi "musnah". Reduksi ini penting untuk mengurangi dimensi fitur sekaligus menghilangkan noise linguistik yang dapat mengganggu proses klasifikasi. Setelah preprocessing, rata-rata panjang teks berkurang signifikan namun tetap mempertahankan kata-kata kunci pembeda antar kategori.

4.3. Hasil Pembobotan TF-IDF

Setelah preprocessing, representasi numerik teks dibentuk menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) dengan konfigurasi 300 fitur dan rentang n-gram (1,2). Tabel 3 menyajikan sepuluh kata dan frasa dengan bobot TF-IDF rata-rata tertinggi dalam keseluruhan corpus.

Tabel 5. Top 10 fitur tf-idf tertinggi

Rank	Kata / Frasa	Bobot TF-IDF
1	sppd	0,2621
2	surat	0,0431
3	sop	0,0379
4	kerja	0,0351
5	tugas monitoring	0,0327
6	surat tugas	0,0327
7	monitoring jaring	0,0327
8	tugas	0,0327
9	monitoring	0,0327
10	jaring	0,0327

Tabel 5 mengungkapkan dua temuan yang menarik. Pertama, kata "sppd" memiliki bobot TF-IDF sebesar 0,2621, jauh melampaui fitur lainnya. Hal ini disebabkan oleh kombinasi dua faktor: frekuensi kemunculannya yang tinggi dalam kategori SPPD, dan sifat kata tersebut yang sangat spesifik sehingga memiliki nilai IDF yang tinggi pula. Ini menjadikan "sppd" sebagai fitur paling diskriminatif dan paling bermanfaat bagi model.

Kedua, penggunaan n-gram (1,2) terbukti efektif dalam menangkap frasa kontekstual seperti "tugas

monitoring", "surat tugas", dan "monitoring jaring". Frasa-frasa tersebut jauh lebih informatif dibandingkan kata tunggalnya karena mencerminkan konteks spesifik administrasi persuratan di lingkungan instansi teknologi informasi. Keputusan menggunakan bigram ini merupakan salah satu kontribusi teknis penting yang membedakan penelitian ini dari pendekatan bag-of-words konvensional.

4.4. Evaluasi Model SVM (Pembagian 80:20)

Model SVM dengan kernel RBF dilatih menggunakan 150 data dan diuji pada 38 data. Hasil evaluasi model ditunjukkan pada Tabel 4 berikut.

Tabel 6. Hasil evaluasi model svm per kategori (80:20)

Kategori	P	R	F1	n
Berita Acara	1,00	0,50	0,67	2
Data Personel	0,00	0,00	0,00	1
Laporan	0,00	0,00	0,00	1
Pengadaan	0,00	0,00	0,00	1
Permohonan	1,00	0,50	0,67	2
Rekomendasi	0,00	0,00	0,00	1
SK	0,36	1,00	0,53	5
SOP	1,00	1,00	1,00	4
SPPD	1,00	1,00	1,00	10
Surat Balasan	1,00	1,00	1,00	1
Surat Kepegawaian	1,00	0,50	0,67	2
Surat Keterangan	0,00	0,00	0,00	1
Surat Pengantar	1,00	1,00	1,00	2
Surat Tugas	1,00	0,80	0,89	5
Undangan	0,00	0,00	0,00	1
Weighted Avg	**0,78**	**0,76**	**0,74**	**38**

Keterangan: P = Precision, R = Recall, F1 = F1-Score, n = jumlah data uji per kelas.

Berdasarkan Tabel 6, model mencapai accuracy 76,32%, precision 78,38%, recall 76,32%, dan F1-score 73,88%. Tiga kategori mencapai nilai sempurna (precision, recall, dan F1-score = 1,00), yaitu SPPD, SOP, dan Surat Pengantar. Ketiganya merupakan jenis surat dengan terminologi yang sangat khas dan konsisten: SPPD selalu mengandung akronim yang sama, SOP selalu diawali kata standar, dan Surat Pengantar memiliki frasa pembuka yang baku.

Surat Tugas juga menunjukkan performa sangat baik dengan F1-score 0,89. Satu kesalahan yang terjadi disebabkan oleh teks surat yang hanya mengandung satu kata kunci tanpa konteks pendukung, sehingga ambigu bagi model. Di sisi lain, kategori SK memiliki precision rendah (0,36) meskipun recall sempurna (1,00), yang berarti model mengenali seluruh SK dengan benar, namun juga salah mengklaim kategori lain sebagai SK. Pola ini akan dibahas lebih lanjut pada analisis confusion matrix.

Nilai nol pada kategori seperti Rekomendasi, Data Personel, Laporan, Surat Keterangan, dan Undangan disebabkan oleh sedikitnya data uji (hanya 1 sampel per kategori). Dengan hanya satu sampel, satu kesalahan prediksi saja sudah cukup untuk menghasilkan nilai nol pada seluruh metrik. Hal ini merupakan keterbatasan yang inheren pada dataset kecil dengan distribusi kelas yang tidak seimbang.

4.5. Hasil 10-Fold Cross Validation

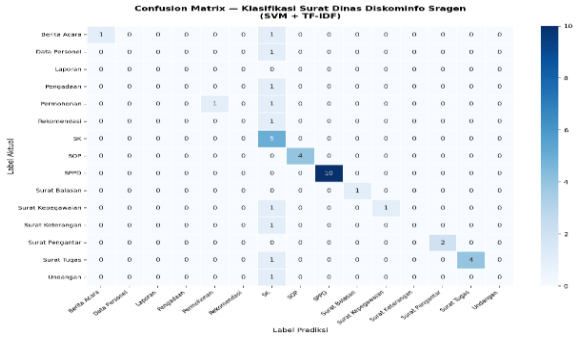
Untuk memperoleh estimasi performa yang lebih stabil dan tidak bergantung pada satu partisi data tertentu, dilakukan evaluasi menggunakan metode 10-fold cross validation. Pada metode ini, data dibagi menjadi 10 bagian dan model dilatih serta diuji sebanyak 10 kali secara bergantian. Hasil selengkapnya disajikan pada Tabel 7.

Tabel 7.Hasil 10-fold cross validation

Fold	Accuracy	Precision	Recall	F1-Score
1	0,7368	0,6382	0,7368	0,6651
2	0,7368	0,6382	0,7368	0,6651
3	0,7895	0,7281	0,7895	0,7333
4	0,6316	0,5737	0,6316	0,5641
5	0,6316	0,5497	0,6316	0,5646
6	0,6842	0,6266	0,6842	0,6281
7	0,6842	0,6266	0,6842	0,6257
8	0,7368	0,6228	0,7368	0,6632
9	0,8333	0,7389	0,8333	0,7731
10	0,8333	0,7500	0,8333	0,7778
Rata-rata	**0,7298**	**0,6493**	**0,7298**	**0,6660**
Std Dev	0,0697	0,0648	0,0697	0,0723

Tabel 7 menunjukkan bahwa model menghasilkan akurasi rata-rata 72,98% dengan standar deviasi 6,97% melalui 10-fold cross validation. Terdapat variasi performa yang cukup jelas antar fold: Fold 4 dan 5 mencatat akurasi terendah (63,16%), sedangkan Fold 9 dan 10 mencapai akurasi tertinggi (83,33%). Perbedaan ini mencerminkan sensitivitas model terhadap komposisi data pada setiap partisi, yang merupakan konsekuensi langsung dari jumlah dataset yang masih terbatas. Nilai standar deviasi sebesar 6,97% masih dalam batas toleransi yang wajar untuk dataset berukuran kecil. Ini mengindikasikan bahwa model cukup konsisten dan tidak mengalami overfitting yang parah. Tren peningkatan performa pada fold 8, 9, dan 10 juga mengisyaratkan bahwa model berpotensi mencapai akurasi yang lebih tinggi apabila jumlah data latih ditambah.

4.6. Analisis Confusion Matrix

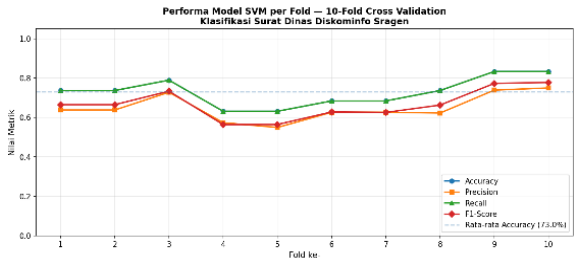


Gambar 2. Confusion matrix klasifikasi surat dinas diskominfo sragen (svm + tf-idf)

Berdasarkan Gambar 2, terdapat pola kesalahan yang dominan, yaitu kecenderungan model mengklasifikasikan berbagai kategori surat ke dalam kelas SK. Dari confusion matrix teridentifikasi setidaknya tujuh kategori berbeda yang memiliki satu atau lebih sampel salah prediksi ke arah SK, antara lain Berita Acara, Permohonan, Surat Kepegawaian, Surat Tugas, Data Personel, Rekomendasi, dan Undangan.

Fenomena ini dapat dijelaskan melalui dua mekanisme. Pertama, SK merupakan kelas terbesar kedua (27 data) setelah SPPD, sehingga model memiliki hyperplane yang lebih condong ke arah SK pada ruang fitur. Kedua, surat bertipe SK sering menggunakan frasa generik seperti "tentang", "nomor", atau nama jabatan, yang juga muncul pada berbagai jenis surat lainnya, sehingga batas keputusan antar kelas menjadi ambigu.

Di sisi lain, keberhasilan sempurna pada kategori SPPD (10/10 prediksi benar), SOP (4/4), dan Surat Pengantar (2/2) membuktikan bahwa SVM sangat efektif pada kelas-kelas yang memiliki ciri linguistik yang kuat dan unik. Ini menunjukkan bahwa kualitas fitur teks jauh lebih berpengaruh terhadap performa model dibandingkan jumlah data semata.



Gambar 3. Grafik performa model svm per fold(10-fold cross validation)

Gambar 3 memperlihatkan tren fluktuasi performa model pada tiap fold. Kurva Recall (hijau) secara konsisten lebih tinggi dari Precision (oranye) dan F1-Score (merah), yang mengindikasikan bahwa model lebih mampu menemukan data yang benar (recall tinggi) dibandingkan menghindari kesalahan prediksi (precision lebih rendah). Ini konsisten dengan pola bias model ke arah kelas SK yang memperbesar recall kelas tersebut namun menurunkan precision secara keseluruhan.

4.7. Pembahasan

Pengujian model klasifikasi menggunakan kombinasi TF-IDF dan SVM menghasilkan akurasi sebesar 76,32% pada skenario pembagian data 80:20. Nilai ini menunjukkan bahwa model mampu melakukan generalisasi yang cukup baik terhadap data uji. Studi sebelumnya menunjukkan bahwa metode TF-IDF mampu menghasilkan representasi fitur yang efektif pada data teks dan secara konsisten memberikan performa yang lebih baik dibandingkan pendekatan sederhana seperti N-Gram[17].

Hal ini menjadikan TF-IDF sebagai pendekatan yang relevan dalam tugas klasifikasi teks berbasis dokumen untuk memvalidasi kestabilan model, dilakukan 10-fold cross-validation yang menghasilkan rata-rata akurasi sebesar 72,98% dengan standar deviasi 6,97%. Perbedaan nilai yang relatif kecil antara skenario hold-out dan cross-validation mengindikasikan bahwa model tidak mengalami overfitting yang signifikan. Metode K-fold cross-validation banyak digunakan dalam evaluasi model karena mampu memberikan estimasi performa yang lebih stabil dengan memanfaatkan seluruh data secara bergantian sebagai data latih dan uji[18].

Nilai standar deviasi yang diperoleh menunjukkan adanya variasi performa antar-fold yang masih dalam batas wajar, yang kemungkinan dipengaruhi oleh distribusi data yang tidak merata.

Analisis per kategori menunjukkan bahwa kelas dengan karakteristik bahasa yang lebih spesifik, seperti SPPD, SOP, dan Surat Pengantar, memiliki tingkat akurasi yang lebih tinggi. Hal ini menunjukkan bahwa TF-IDF bekerja lebih optimal ketika terdapat perbedaan kosakata yang jelas antar kategori. Penelitian sebelumnya juga menyatakan bahwa efektivitas TF-IDF sangat bergantung pada tingkat perbedaan fitur antar kelas, di mana representasi berbasis frekuensi kata lebih unggul pada data dengan karakteristik linguistik yang distingtif (Li et al., 2023). Sebaliknya, kategori dengan kemiripan kosakata yang tinggi cenderung menghasilkan performa yang lebih rendah karena keterbatasan TF-IDF dalam memahami konteks semantik.

Selain itu, ketidakseimbangan distribusi kelas (class imbalance) juga menjadi faktor penting yang memengaruhi performa model. Kondisi ini menyebabkan model cenderung lebih fokus pada kelas mayoritas selama proses pembelajaran, sehingga kemampuan dalam mengenali kelas minoritas menjadi menurun. Akibatnya, meskipun nilai akurasi terlihat tinggi, performa model pada kelas tertentu tidak optimal[19].

Beberapa penelitian menunjukkan bahwa pada kondisi data tidak seimbang, metrik seperti F1-Score lebih representatif dibandingkan akurasi karena mempertimbangkan keseimbangan antara precision dan recall [20].

Oleh karena itu, penggunaan macro-averaged F1-Score dalam penelitian ini dinilai lebih tepat untuk menggambarkan performa model secara keseluruhan.

Selain jumlah data, kualitas distribusi fitur juga berpengaruh terhadap performa klasifikasi. Penelitian sebelumnya menunjukkan bahwa kesamaan fitur dalam satu kelas (intra-class similarity) dan perbedaan fitur antar kelas (inter-class dissimilarity) memiliki peran penting dalam menentukan keberhasilan model klasifikasi [21].

Hal ini sejalan dengan kondisi pada dataset surat dinas, di mana beberapa kategori memiliki struktur bahasa yang mirip sehingga menyulitkan model dalam membedakan kelas secara optimal.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi TF-IDF dan SVM

merupakan pendekatan yang efektif dalam klasifikasi dokumen administratif berbasis teks pendek. Temuan ini juga didukung oleh berbagai penelitian yang menunjukkan bahwa SVM memiliki kemampuan yang baik dalam menangani data berdimensi tinggi serta menghasilkan model klasifikasi yang stabil [22].

Untuk pengembangan selanjutnya, penerapan teknik penyeimbangan data seperti SMOTE serta penggunaan model berbasis representasi kontekstual seperti Transformer atau IndoBERT dapat menjadi solusi untuk meningkatkan performa dan mengatasi keterbatasan metode saat ini.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian metode text mining menggunakan pembobotan TF-IDF dan algoritma Support Vector Machine (SVM) terbukti cukup efektif untuk mengklasifikasikan surat dinas. Hal ini dibuktikan dengan akurasi 76,32% pada pengujian rasio 80:20, serta rata-rata 72,98% pada 10-fold cross validation. Pendekatan ini berhasil meningkatkan efisiensi, konsistensi, dan kecepatan pengelolaan surat dinas dibandingkan metode manual. Meski demikian, performa model masih terhambat oleh ketidakseimbangan data dan ukuran dataset yang terbatas. Untuk itu, penelitian lanjutan disarankan menggunakan dataset lebih besar dan seimbang, serta mencoba metode deep learning atau optimasi parameter guna meningkatkan hasil model.

DAFTAR PUSTAKA

[1] L. Jaillant and L. Zhao, "Introduction : When data turns into archives : making digital records more accessible with AI," *AI Soc.*, no. 123, 2025.

[2] N. Fajri, A. Rizaldy, and A. M. Abidin, "Penerapan Natural Language Processing Menggunakan TF-IDF dan N-Gram untuk Layanan Donor Darah," *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 6, no. 1, 2026.

[3] A. D. M. Putri, N. Sulistianingsih, and R. Rismayati, "Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 7, no. 4, pp. 675–688, 2025.

[4] K. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines : Principles , Techniques , and Future Directions," *Mathematics*, vol. 12, no. 24, p. 3935, 2024, doi: 10.3390/math12243935.

[5] A. H. Al-Habib, E. M. Imah, R. D. I. Puspitasari, and B. K. Prahani, "Text Processing Using Support Vector Machine for Scientific Research Paper Content Classification," in *International Joint Conference on Science and Engineering 2022 (IJCSE 2022)*, Atlantis Press International BV, 2023, pp. 273–282.

[6] K. Piasta and R. Kotas, "Comparative Analysis of Natural Language Processing Techniques in the Classification of Press Articles," *Appl. Sci.*, vol. 15, no. 11, p. 5595, 2025.

[7] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset," in *CEUR Workshop Proceedings*, 2021.

[8] C.-A. Tsai and Y.-J. Chang, "Efficient Selection of Gaussian Kernel SVM Parameters for Imbalanced Data," *Genes (Basel).*, vol. 14, no. 3, p. 583, 2023.

[9] K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Sci. J.*, vol. 23, no. 31, 2024.

[10] Istiadi, A. Y. Rahman, and K. S. Nugroho, "A Comparative Study of Text Classification Performance Using NBC and KNN with N-Gram Features in E-Government Services," in *Proceedings of the 1st International Conference on Business, Innovation, Technology & Science (ICoBITS)*, 2024, pp. 1006–1012.

[11] Abdurrazika and I. M. W. Wirawana, "Analisis Klasifikasi Tweet Berdasarkan Topik Sosial Menggunakan SVM," *JNATIA*, vol. 3, no. 4, pp. 855–864, 2025.

[12] T. Abedin, H. Xu, and S. Uddin, "The impact of K selection in K - fold cross-validation on bias and variance in supervised learning models," *Sci. Rep.*, vol. 16, pp. 1–13, 2026.

[13] O. Chamorro-Atalaya *et al.*, "K-Fold Cross-Validation through Identification of the Opinion Classification Algorithm for the Satisfaction of University Students," *Int. J. Online Biomed. Eng.*, vol. 19, no. 11, pp. 140–158, 2023.

[14] D. Canning and L. Jaillant, "MAIN PAPER AI to review government records : new work to unlock historically significant digital records," *AI Soc.*, vol. 40, no. 6, pp. 4433–4445, 2025.

[15] K. Taha, P. D. Yoo, C. Yeun, and A. Taha, "Text Classification Techniques : A Holistic Review , Observational Analysis , and Experimental Investigation," vol. 8, no. 3, pp. 624–660, 2025.

[16] B. Santoso, *Support Vector Machine (Algoritma dan Aplikasinya)*. Padang: CV. Karya Buku dan Jurnal Indonesia, 2025.

[17] K. Taha, P. D. Yoo, C. Yeun, D. Houmouz, and A. Taha, "A Comprehensive Survey of Text Classification Techniques and Their Research Applications: Observational and Experimental Insights," *arXiv Prepr.*, vol. 8, 2024.

[18] D. Pakpahan, V. Siallagan, and S. Siregar, "Classification of E-Commerce Product Descriptions with The Tf-Idf and Svm Methods," *Sinkron*, vol. 8, no. 4, pp. 2130–2137, 2023.

[19] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "OPEN A comprehensive evaluation of oversampling techniques for enhancing text classification performance," pp. 1–20, 2025.

[20] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P.

Chen, *A survey on imbalanced learning : latest research , applications and future directions*, vol. 6. 2024.

[21] S. Feng, P. Jin, and C. Wang, "CASE: Exploiting Intra-class Compactness and Inter-class Separability of Feature Embeddings for Out-of-Distribution Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 21081–21089.

[22] A. M. Aubaid, A. Mishra, and A. Mishra, "Machine learning and rule - based embedding techniques for classifying text documents," *Int. J. Syst. Assur. Eng. Manag.*, vol. 15, no. 12, pp. 5637–5652, 2024.