KLASIFIKASI STATUS PEMERIKSAAN DIGITALISASI KOLEKSI PERPUSTAKAAN BADAN PUSAT STATISTIK REPUBLIK INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE LINEAR*

Fathinah Nur Azizah[1], Luluk Muthoharoh[2*], Wiwik Akta Piantri[3]
[1,2]Sains Data, Institut Teknologi Sumatera, Bandar Lampung, Indonesia
[3]BPS Republik Indonesia

**Penulis korespondensi: luluk.muthoharoh@sd.itera.ac.id*

ABSTRAK

Perpustakaan Badan Pusat Statistik Republik Indonesia (BPS RI) berperan sebagai pusat informasi yang menunjang kegiatan akademik sekaligus menyediakan data dan literatur ilmiah. Proses *quality control* dalam kegiatan digitalisasi koleksi perpustakaan selama ini dilakukan secara manual, sehingga rentan terhadap inkonsistensi dan membutuhkan waktu yang tidak sedikit. Penelitian ini bertujuan untuk mengklasifikasikan status pemeriksaan digitalisasi koleksi perpustakaan BPS RI secara otomatis menggunakan metode Support Vector Machine (SVM) Linear dengan pendekatan *One-vs-Rest*. Data yang digunakan merupakan catatan hasil *quality control* periode Agustus hingga Desember 2025 yang terdiri dari 2657 baris dan 9 kolom. Data tersebut bersifat tidak terstruktur sehingga memerlukan serangkaian proses *preprocessing* sebelum dianalisis lebih lanjut. Tahapan pemodelan mencakup *preprocessing* teks, transformasi fitur menggunakan TF-IDF, serta pelatihan model klasifikasi untuk memprediksi tiga kategori status, yaitu "Baik", "Perlu Perbaikan", dan "Tidak Ada Status". Hasil evaluasi model menunjukkan akurasi sebesar 93%, presisi 91%, *recall* 77%, dan *F1-Score* 81%, yang mengindikasikan bahwa model mampu mengklasifikasikan status digitalisasi dengan baik meskipun terdapat ketidakseimbangan antar kelas. Model yang dihasilkan diharapkan dapat memberikan rekomendasi status digitalisasi secara otomatis, sehingga meningkatkan efisiensi, konsistensi, dan objektivitas proses *quality control* di Perpustakaan BPS RI.

Kata kunci: digitalisasi koleksi perpustakaan, klasifikasi teks, *quality control*, *support vector machine*, TF-IDF

1 PENDAHULUAN

Perpustakaan Badan Pusat Statistik Republik Indonesia (BPS RI) merupakan salah satu perpustakaan dengan peran penting dalam menyediakan sumber informasi yang dibutuhkan pengunjung serta berperan penting dalam mendukung kegiatan akademik berupa penelitian dan membantu dalam meningkatkan reputasi BPS RI (Utami dan Sihite, 2021). Manajemen mutu Perpustakaan Badan Pusat Statistik Indonesia meliputi kegiatan, relokasi, alih media (digitalisasi), pemeriksaan hasil, dan publikasi. Proses pemeriksaan hasil perlu diperhatikan karena mampu berpengaruh terhadap tingkat kepuasan pengunjung karena memiliki hubungan dengan sumber informasi yang dapat diakses serta dapat membantu proses evaluasi layanan perpustakaan (Nero dan He, 2018; Puspita dkk., 2024). Pada penerapannya di BPS RI hasil pemeriksaan hasil akan menjadi penentu sebuah publikasi buku yang berbentuk klasifikasi apakah buku sudah layak dipublikasi atau perlu digitalisasi kembali.

Data hasil pemeriksaan hasil merupakan data teks yang berisikan catatan hasil pemeriksaan hasil yang dilakukan oleh Tim Perpustakaan BPS RI serta klasifikasi secara manual untuk menentukan keputusan publikasi koleksi perpustakaan. Proses manual ini berisiko terhadap pelabelan yang tidak konsisten sehingga pemodelan menggunakan *machine learning*

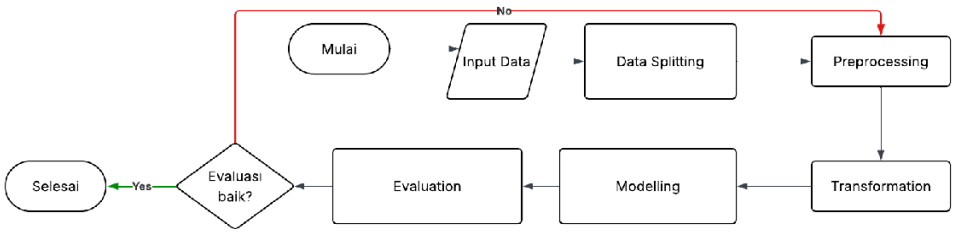
diharapkan mampu membantu permasalahan tersebut. Hal ini dapat diatasi menggunakan *text mining* yang merupakan salah satu ilmu *data mining* yang bertujuan untuk mendapatkan informasi penting dari sebuah teks menggunakan pembobotan kata (Anggraini dkk., 2021).

Data yang diambil berupa teks yaitu data catatan yang dilakukan data *preprocessing* dan transformasi menggunakan *Term Frequency Inverse Document Frequency* (TF-IDF) yang akan memberikan bobot pada kata-kata berdasarkan frekuensinya dalam sebuah dokumen dan frekuensi terbaliknya di seluruh korpus, dengan cara menekankan istilah umum dalam dokumen tertentu tapi *relative* jarang di seluruh kumpulan data (Occhipinti dkk., 2022) Penelitian lain menyebutkan bahwa algoritma *Support Vector Machine* (SVM) yang merupakan salah satu metode regresi atau klasifikasi data berdasarkan data terdahulu memiliki akurasi terbaik (Deolika dkk., 2019). Klasifikasi atau pembagian data SVM dibantu dengan kernel yang secara default dibantu oleh 4 parameter kernel yaitu *linear, polynomial, radial basis function,* dan *sigmoid* (Riadi dkk., 2019).

Berdasarkan kondisi tersebut, penelitian ini berfokus pada penerapan metode *machine learning* untuk mengklasifikasikan status hasil pemeriksaan hasil digitalisasi berdasarkan data teks hasil pemeriksaan digitalisasi koleksi Perpustakaan BPS RI. Penelitian ini akan mengolah data teks menggunakan pembobotan TF-IDF unigram (satu kata) dan bigram (dua kata) serta diklasifikasikan dengan algoritma *Linear Support Vector Machine* (SVM Linear). Melalui pendekatan ini, diharapkan dapat diperoleh model yang mampu membantu proses penentuan status hasil pemeriksaan hasil digitalisasi secara otomatis dan memberikan gambaran performa metode *machine learning* dalam menangani permasalahan klasifikasi pada data pemeriksaan hasil dokumen digital.

2 METODE

Metode yang digunakan pada penelitian ini dijabarkan pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

2.1 *Input* Data

Data yang digunakan untuk penelitian merupakan data hasil proses pemeriksaan hasil tim perpustakaan dalam memeriksa hasil digitalisasi koleksi Perpustakaan BPS RI. *Dataset* berisi 2657 baris dengan 9 atribut. Fitur yang tersedia antara lain bulan, lokasi, nama box, nomor box, nomor eksemplar, judul, jumlah halaman, dan catatan serta label target status.

Tabel 1. Data Mentah

Bulan	Lokasi	Nama_Box	…	Judul	Jumlah_Halaman	Catatan	Status
Agustus	Box	SCN	…	STATUS GIZI BATITA-- HASIL INTEGRASI ANTROPOMETRI DALAM SUSENAS 1985/1986	125	-	Baik
Agustus	Box	SCN	…	ANALISIS TINGKAT KESEJAHTERAAN PENDUDUK DAN	76	sesuai fisik buku ada	Baik

Bulan	Lokasi	Nama_Box	…	Judul	Jumlah_Halaman	Catatan	Status
				PERKEMBANGANNYA BERDASARKAN DATA SENSUS PENDUDUK 1980 DAN SURVEI PENDUDUK ANTAR SENSUS 1985		hal tidak ada (hilang): 28	
Agustus	Box	SCN	…	KABUPATEN SUMBAWA BARAT DALAM ANGKA 2011	425	Bab XIII (hal buku 445-451) tidak ada. di daftar isi ada hal tersebut. apakah di fisik bukunya memang tidak ada?	Perlu Perbaikan
Desember	BOX	SCN	…	STATISTIK SOSIAL DAN KEPENDUDUKAN NUSA TENGGARA TIMUR 2003	157	-	Baik
Desember	BOX	SCN	…	STATISTIK SOSIAL DAN KEPENDUDUKAN NUSA TENGGARA TIMUR 2002	148	-	Baik
Desember	BOX	SCN	...	PENDUDUK NUSA TENGGARA BARAT HASIL SENSUS PENDUDUK 1980	183	-	Baik

Tabel 1 menunjukkan bahwa mayoritas data memiliki tipe *character* atau teks, sehingga fokus penelitian ini terhadap *text mining* adalah menjadikan atribut catatan sebagai fitur, sedangkan atribut status sebagai target.

2.2 Data *Splitting*

Penelitian ini menggunakan *dataset* pemeriksaan hasil digitalisasi koleksi buku Perpustakaan BPS RI yang memiliki atribut catatan sebagai penentu status hasil digitalisasi sehingga pada tahap *data splitting*, catatan akan menjadi variabel X, sedangkan atribut status akan menjadi variabel Y. *Data splitting* terbagi menjadi dua yaitu *data training* sebagian data latih dalam membangun model dan *data testing* sebagai data uji terhadap model. Setelah itu, kedua atribut dibagi menjadi *data training* dan *data testing* dengan rasio 80:20 dengan fungsi *train_test_split* pustaka ‘sklearn.model_selection’ dengan parameter *test_size* 0.2 yang berarti 20% data akan digunakan untuk data uji sehingga 80% sisanya menjadi data latih.

2.3 *Preprocessing*

Preprocessing merupakan tahapan pengolahan data yang dapat meningkatkan hasil transformasi dan *modelling* yang baik. *Preprocessing* memiliki beberapa tahap yaitu *data cleaning, case folding,* dan *tokenizing* tanpa *stopword removal* maupun *stemming*. Keputusan

ini berdasarkan pada karakteristik teks yang digunakan, yaitu catatan pemeriksaan yang bersifat teknis sehingga setiap katanya memiliki makna kontekstual yang mempengaruhi status digitalisasi dokumen. Jika terdapat *stopword removal* maka berpotensi menghilangkan makna dari tiap catatannya. Selain itu, *stemming* tidak digunakan karena akan menghilangkan makna inti dari tiap data catatan.

2.3.1 Data *Cleaning*

Data cleaning merupakan tahapan untuk membersihkan data teks dari symbol seperti, !, @, #, $, %, ^, &, *, (,), <,>, dan simbol serta angka lainnya yang akan dihapus (Sumitro dkk., 2021).

2.3.2 *Case Folding*

Case folding merupakan tahapan untuk mengubah semua huruf menjadi huruf kecil (Salam dkk., 2023).

2.3.3 *Tokenizing*

Tokenizing merupakan tahapan untuk memecah kalimat menjadi kata per kata berdasarkan spasi (Suyati dan Aldino, 2023).

2.4 *Transformastion*

Transformation dilakukan untuk mendapatkan bobot pada setiap katanya sehingga memudahkan proses *data mining* (Herwijayanti dkk, 2018). Pembobotan dilakukan menggunakan *Term Frequency Inverse Document Frequency* (TF-IDF) yang akan menentukan hubungan kata (*term*) terhadap kalimat dengan dilakukan perhitungan terlebih dahulu nilai TF (Arifin dkk., 2021). Pembobotan TF-IDF pada penelitian ini diterapkan pada teks yang pernah melalui tahap *preprocessing*. Konfigurasi TF-IDF menggunakan kombinasi fitur *unigram* dan *bigram* (ngram_range=(1,2)) melalui pustaka 'TfidfVectorized' untuk menangkap konteks frasa pendek yang tidak tertangkap oleh *unigram*. Parameter *max_features* dibatasi 1000 fitur secara keseluruhan, dipilih berdasarkan bobot TF-IDF tertinggi dari setiap kata. Selain itu, parameter *min_df* dan *max_df* tidak diatur sehingga seluruh token yang muncul minimal satu kali tetap dipertimbangkan sebelum disaring oleh *max_features*. Pembatasan jumlah fitur bertujuan untuk mengurangi *overfitting*.

2.5 *Modelling*

Modelling merupakan tahapan dalam *machine learning* dalam pembuatan model menggunakan data *training*. Penelitian ini menggunakan SVM Linear yang akan melibatkan batas keputusan berupa *hyperplane* yang memisahkan 3 kelas (Borman dan Wati, 2020). Adanya 3 kelas maka model akan menggunakan prinsip *One vs Rest* (OVR) (Psaltakis dkk., 2024). Model SVM Linear pada penelitian ini dibangun menggunakan pustaka LinearSVC dari *scikit-learn* dengan parameter regulasi C=1.0 (nilai *default*), yang mengatur keseimbangan antara memaksimalkan *margin* dan meminimalkan kesalahan klasifikasi pada data *training*. Pemilihan nilai *default* ini menjadi salah satu keterbatasan penelitian, mengingat penyesuaian nilai C melalui proses *tuning* (misalnya menggunakan *grid search*) berpotensi meningkatkan kemampuan generalisasi model, khususnya pada kelas dengan jumlah data yang lebih sedikit.

2.6 *Evaluation*

Evaluasi merupakan tahap akhir untuk menampilkan tingkat akurasi dan ketetapan dari hasil klasifikasi yang direpresentasi sebagai *confusion matrix* yang berfungsi untuk mengevaluasi model berdasarkan kelas yang ada (Rindiyani dkk., 2022). Komponen *confusion*

matrix penelitian ini adalah indeks *(i, j)* di mana *i* sebagai indeks prediksi sedangkan *j* indeks *actual*.

2.6.1 *Precision*

Adanya tiga kelas dalam data ini maka persamaan *precision* memiliki 3 persamaan yaitu:

$$Precision_{Baik} = \frac{T_{Baik}}{T_{Baik}+F_{Baik,Perlu\ Perbaikan}+F_{Baik,Tidak\ Ada\ Status}} \quad (1)$$

$$Precsision_{Perlu\ Perbaikan} = \frac{T_{Perlu\ Perbaikan}}{T_{Perlu\ Perbaikan}+F_{Perlu\ Perbaikan,Baik}+F_{Perlu\ Perbaikan,Tidak\ Ada\ Status}} \quad (2)$$

$$Precision_{Tidak\ Ada\ Status} = \frac{T_{Tidak\ Ada\ Status}}{T_{Tidak\ Ada\ Status}+F_{Tidak\ Ada\ Status,Baik}+F_{Tidak\ Ada\ Status,Perlu\ Perbaikan}} \quad (3)$$

2.6.2 *Recall*

Recall merupakan perbandingan data kelas sesuai dengan data kelas sesuai dan data diprediksi salah kelas aslinya sesuai kelas. Dengan persamaan matematis untuk tiap kelasnya adalah:

$$Recall_{Baik} = \frac{T_{Baik}}{T_{Baik}+F_{Baik,Perlu\ Perbaikan}+F_{Baik,Tidak\ Ada\ Status}} \quad (4)$$

$$Recall_{Perlu\ Perbaikan} = \frac{T_{Perlu\ Perbaikan}}{T_{Perlu\ Perbaikan}+F_{Perlu\ Perbaikan,Baik}+F_{Perlu\ Perbaikan,Tidak\ Ada\ Status}} \quad (5)$$

$$Recall_{Tidak\ Ada\ Status} = \frac{T_{Tidak\ Ada\ Status}}{T_{Tidak\ Ada\ Status}+F_{Tidak\ Ada\ Status,Baik}+F_{Tidak\ Ada\ Staus,Perlu\ Perbaikan}} \quad (6)$$

2.6.3 *Accuracy*

Accuracy merupakan perbandingan dari total prediksi benar dengan total data dengan persamaan matematis:

$$Accuracy = \frac{T_{Baik}+T_{Perlu\ Perbaikan}+T_{Tidak\ Ada\ Status}}{Total\ Data} \quad (7)$$

2.6.4 *F1-Score*

F1-Score merupakan perhitungan dengan membandingkan perkalian *precision* dan *recall* dengan penjumlahan *precision* dan *recall.* Dengan persamaan matematis untuk tiap kelasnya adalah:

$$F1-Score_{Baik} = 2 \times \frac{Precision_{Baik} \times Recall_{Baik}}{Precision_{Baik}+Recall_{Baik}} \quad (8)$$

$$F1-Score_{Perlu\ Perbaikan} = 2 \times \frac{Precision_{Perlu\ Perbaikan} \times Recall_{Perlu\ Perbaikan}}{Precision_{Perlu\ Perbaikan}+Recall_{Perlu\ Perbaikan}} \quad (9)$$

$$F1-Score_{Tidak\ Ada\ Status} = 2 \times \frac{Precision_{Tidak\ Ada\ Status} \times Recall_{Tidak\ Ada\ Status}}{Precision_{Tidak\ Ada\ Status}+Recall_{Tidak\ Ada\ Status}} \quad (10)$$

3 HASIL DAN PEMBAHASAN

3.1 Hasil *Data Preprocessing*

Data pada penelitian ini memiliki permasalahan tidak ada kelas dan ada beberapa data kosong pada fitur catatan. Kelas yang dimiliki yaitu kelas "Baik" sebanyak 2087 data, "Tidak Ada Status" sebanyak 114 data, dan "Perlu Perbaikan" sebanyak 456 data. Proses *data preprocessing* dilakukan dengan tujuan penyeragaman terhadap data teks dengan rincian proses berupa *data cleaning*, *case folding*, dan *tokenizing*.

Tabel 2. Perbandingan Sebelum dan Sesudah *Data Cleaning* pada Fitur Catatan

Sebelum *Cleaning*	Sesudah *Cleaning*
	Sudah Baik
sesuai fisik buku ada hal tidak ada (hilang)	sesuai fisik buku ada hal tidak ada hilang
28Bab XIII (hal buku 445-451) tidak ada.	Bab XIII hal buku tidak ada.
di daftar isi ada hal tersebut. apakah di fisik bukunya memang tidak ada?	daftar isi ada hal tersebut. apakah fisik bukunya memang tidak ada
…	…
	Sudah Baik
	Sudah Baik

Tahapan *data cleaning* juga meliputi penanganan data kosong pada fitur catatan dengan menambahkan tulisan "Sudah baik", karena semua data kosong pada fitur catatan memiliki label status "Baik". Sehingga, pengisian ini harapannya dapat membantu algoritma dalam membangun model serta penghapusan simbol dan angka yang dibantu oleh pustaka 're' dalam membersihkan data teks.

Tabel 3. Perbandingan Sebelum dan Sesudah *Case Folding* pada Fitur Catatan

Sebelum *Case Folding*	Sesudah *Case Folding*
Sudah Baik	sudah baik
sesuai fisik buku ada hal tidak ada (hilang)	sesuai fisik buku ada hal tidak ada(hilang
Bab XIII (hal buku) tidak ada.	bab XIII hal buku tidak ada.
daftar isi ada hal tersebut. apakah fisik bukunya memang tidak ada	daftar isi ada hal tersebut. apakah fisik bukunya memang tidak ada
…	…
Sudah Baik	sudah baik
Sudah Baik	sudah baik

Case folding dilakukan untuk mengubah data teks yang berhuruf besar menjadi data teks berhuruf kecil sehingga terdapat keseragamann pengaturan huruf.

Tabel 4. Perbandingan Sebelum dan Sesudah *Tokenizing* pada Fitur Catatan

Sebelum *Tokenizing*	Sesudah *Tokenizing*
sudah baik	[“sudah”, “baik”]
sesuai fisik buku ada hal tidak ada hilang	[“sesuai”, “fisik”, “buku”,“ada”, “hal”, “tidak”, “ada”, “hilang”]
bab XIII hal buku tidak ada. daftar isi ada hal tersebut. apakah fisik bukunya memang tidak ada	[“bab”, ”XIII”, ”hal”, ”buku”, ”tidak”, “ada”, ”daftar”, ”isi”, ”ada”, ”hal”, ”tersebut”, “apakah”, ”fisik”, ”bukunya”, ”memang”, ”tidak”, ”ada”]
…	…
sudah baik	[“sudah”, “baik”]
sudah baik	[“sudah”, “baik”]

Terakhir, penyempurnaan data dilakukan menggunakan *tokenizing* untuk memisahkan kalimat dalam satu catatan menjadi kata per kata sehingga terlihat bentuk dasarnya (Kalaivani dan Marivendan, 2021). *Data cleaning, case folding,* dan *tokenizing* perlu dilakukan dalam proses *text mining* karena dapat meningkatkan kualitas data sebelum dilakukan proses TF-IDF.

3.2 Hasil Transformasi

Hasil transformasi menggunakan TF-IDF dibantu dengan pustaka ‘TfidVectorized’ di dalam *python* untuk melakukan pembobotan setiap komponen variabel *x* terhadap variabel *y*.

Tabel 5. Hasil Pembobotan TF-IDF 5 Kata

	TF				TF-IDF		
	Dokumen 1	Dokumen 2	Dokumen 3	IDF	Dokumen 1	Dokumen 2	Dokumen 3
pdf rotate	0	0	0	7.0525	0	0	0
sesuai	0	0	0	3.917	0	0	0
hal buku	0	0	0	4.9243	0	0	0
halaman	0	0	0	3.898	0	0	0
ada folder	0	0	0	5.828	0	0	0

Hasil perhitungan TF-IDF menunjukkan bahwa beberapa kata memiliki nilai TF dan TF-IDF sebesar nol pada dokumen yang dianalisis, meskipun nilai IDF untuk kata tersebut tidak bernilai nol. Kondisi ini terjadi karena nilai IDF dihitung berdasarkan keseluruhan data pelatihan yang digunakan untuk membangun model, sehingga setiap kata yang pernah muncul dalam korpus akan memiliki nilai IDF tertentu sesuai dengan tingkat kemunculannya. Namun, nilai TF bergantung pada frekuensi kemunculan kata dalam dokumen tertentu yang sedang dianalisis. Apabila suatu kata tidak muncul dalam dokumen tersebut, maka nilai TF akan bernilai nol. Karena TF-IDF merupakan hasil perkalian antara TF dan IDF, maka ketika nilai TF sama dengan nol, nilai TF-IDF juga akan bernilai nol, terlepas dari besarnya nilai IDF. Hal ini merupakan kondisi yang wajar dalam analisis teks, terutama ketika jumlah dokumen yang diamati terbatas atau ketika kata yang dipilih tidak terdapat dalam dokumen tertentu, sehingga tidak memberikan kontribusi terhadap representasi fitur pada dokumen tersebut. Sehingga 5 kata di atas merupakan kata-kata yang penting dalam klasifikasi berdasarkan kumpulan catatan (Yao dkk., 2019).

3.3 Hasil Modelling

Hasil *modelling* menghasilkan persamaan matematis berdasarkan pembobotan yang telah dilakukan melalui transformasi menggunakan algoritma SVM Linear yang dibantu oleh pustaka 'LinearSVC'. Algoritma SVM Linear akan mencari *hyperplane* atau bidang pemisah untuk memisahkan ketiga kelas secara maksimal (Safitri dkk., 2023). Adanya *hyperplane* ini menghasilkan 3 persamaan garis sesuai label target sebagai berikut:

Tabel 6. Model *Hyperplane*

Label Target	Model *Hyperplane*	
"Baik"	$f_{"Baik"}(x_1, x_2, x_3) = -0.2356 + 0.1823x_1 - 0.0279x_2 - 0.5180x_3$	(11)
"Perlu Perbaikan"	$f_{"Perlu\ Perbaikan"}(x_1, x_2, x_3) = 0.0907 - 0.1696x_1 - 0.4003x_2 + 0.7553x_3$	(12)
"Tidak Ada Status"	$f_{"Tidak\ Ada\ Status"}(x_1, x_2, x_3) = -0.8198 - 0.0015 + 0.4067x_2 - 0.2549x_3$	(13)

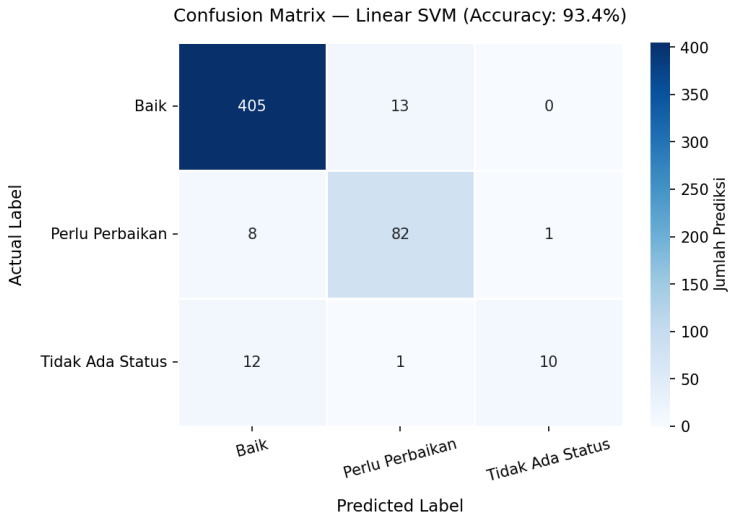
Hasil pada Tabel 6 di atas menunjukkan bahwa persamaan garis tersebut merupakan persamaan garis yang terbaik dalam memisahkan data antar kelas. Dengan menggunakan algoritma SVM Linear dan metode OVR, model menghasilkan 1000 fitur untuk setiap kelasnya. Adanya persamaan matematis ini menunjukkan bahwa SVM Linear didefinisikan secara terpisah setiap kelas menggunakan kombinasi dari koefisien fitur, dan *intercept*. Hal ini diperkuat karena SVM Linear tidak melakukan transformasi lebih lanjut ke data dengan dimensi yang lebih tinggi sehingga *hyperplane* langsung dianggap optimal dalam memisahkan data pada ruang fitur asli (Fitriyah dkk., 2020).

3.4 Hasil Evaluasi

Setelah data berhasil mendapatkan model berdasarkan *data training* maka dilakukan uji coba menggunakan *data testing* yang kemudian dibandingkan untuk menghitung akurasi, *recall*, presisi, dan *F1-Score* menggunakan pustaka '*classification_report*'.

Tabel 7. Hasil Evaluasi

Kelas	*Akurasi*	*Presisi*	*Recall*	*F1-Score*
"Baik"	0.93	0.95	0.97	0.96
"Perlu Perbaikan"		0.85	0.90	0.88
"Tidak Ada Status"		0.91	0.43	0.59
Rata-rata		0.91	0.77	0.81



Gambar 2. *Confusion Matrix*

Meskipun performa keseluruhan tergolong baik, terdapat kesenjangan yang cukup signifikan pada metrik *recall* antar kelas, di mana kelas "Baik" mencapai *recall* 0.97 sedangkan

kelas "Tidak Ada Status" hanya mencapai 0.43. Rendahnya *recall* pada kelas "Tidak Ada Status" diduga disebabkan oleh dua faktor utama. Pertama, adanya ketidakseimbangan jumlah data antar kelas pada data latih, di mana kelas dengan jumlah sampel yang lebih sedikit cenderung lebih sulit dikenali oleh model karena algoritma cenderung memprediksi kelas mayoritas untuk meminimalkan kesalahan klasifikasi secara agregat (Sun dkk., 2009; Leevy dkk., 2018). Kedua, kemiripan karakteristik tekstual antara kelas "Tidak Ada Status" dan kelas "Baik", keduanya sama-sama memiliki catatan yang kosong atau sangat singkat. Hal tersebut menyebabkan representasi TF-IDF kedua kelas tersebut saling tumpang tindih dalam ruang fitur, sehingga *hyperplane* SVM kesulitan memisahkan keduanya secara tegas. Pola serupa juga ditemukan pada penelitian klasifikasi sentimen berbasis SVM lainnya, di mana kelas dengan representasi data paling sedikit secara konsisten menghasilkan *recall* yang jauh lebih rendah dibandingkan kelas mayoritas (Wiguna dkk., 2025).

Hasil evaluasi menunjukkan bahwa prediksi menggunakan SVM Linear dalam menebak kelas yang sesuai mencapai 93% dengan rata-rata 77%, rata-rata presisi 91%, dan rata-rata *F1-Score* 81%. Hasil ini cenderung baik yang diperkuat oleh penelitian terhadulu dalam membandingkan performa kernel pada model SVM yang menunjukkan bahwa kernel linear, RBF, dan *Polynomial* memiliki hasil yang tidak jauh berbeda dengan akurasi 87% untuk setiap kernelnya (Riza, 2022). Pada penelitian lain dengan topik *mental health* yang melibatkan data teks menunjukkan bahwa SVM Linear memiliki akurasi yang cenderung mirip dengan RBF, *polynom,* dan sigmoid yaitu 92% (Minatika dan Arifudin, 2025). Hal ini menunjukkan bahwa SVM Linear sesuai dan cukup baik untuk melakukan *modelling* data teks dengan bantuan proses sebelum tahap *modelling*.

Hasil evaluasi menunjukkan bahwa prediksi menggunakan SVM Linear dalam menebak kelas yang sesuai mencapai akurasi 93%, dengan rata-rata presisi 91%, rata-rata *recall* 77%, dan rata-rata *F1-Score* 81%. Hasil evaluasi menunjukkan bahwa prediksi menggunakan SVM Linear dalam menebak kelas yang sesuai mencapai akurasi 93% dengan rata-rata *77%*, rata-rata presisi 91%, dan rata-rata *F1-Score* 81%. Hasil ini cenderung baik yang diperkuat oleh penelitian terhadulu dalam membandingkan performa kernel pada model SVM yang menunjukkan bahwa kernel linear, RBF, dan *Polynomial* memiliki hasil yang tidak jauh berbeda dengan akurasi 87% untuk setiap kernelnya (Riza, 2022). Pada penelitian lain dengan topik *mental health* yang melibatkan data teks menunjukkan bahwa SVM Linear memiliki akurasi yang cenderung mirip dengan RBF, *polynomial,* dan sigmoid yaitu 92% (Minatika dan Arifudin, 2025). Hal ini menunjukkan bahwa SVM Linear sesuai dan cukup baik untuk melakukan *modelling* data teks dengan bantuan proses sebelum tahap *modelling.* Temuan ini menegaskan bahwa SVM Linear cukup sesuai untuk pemodelan data teks, namun penanganan ketidakseimbangan kelas, misalnya melalui teknik *class weighting* atau *oversampling* seperti SMOTE dan perlu dipertimbangkan pada penelitian selanjutnya untuk meningkatkan *recall* pada kelas minoritas.

4 KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan, dapat disimpulkan bahwa proses *preprocessing* data mampu menyeragamkan data, dan model Linear SVM yang dibantu TF-IDF mampu mengklasifikasikan status digitalisasi dengan akurasi 93%. Meskipun demikian, performa model belum merata antar kelas, ditunjukkan oleh *recall* kelas "Tidak Ada Status" (43%) yang jauh lebih rendah dibandingkan kelas "Baik" (97%) dan "Perlu Perbaikan" (90%), kemungkinan akibat ketidakseimbangan jumlah data antar kelas serta keterbatasan periode data (Agustus–Desember 2025) dan parameter model yang masih menggunakan nilai *default*.

Penelitian selanjutnya disarankan untuk membandingkan Linear SVM dengan model lain seperti *Naive Bayes* dan *Logistic Regression,* menguji berbagai jenis kernel SVM (linear, *polynomial*, RBF), serta menerapkan teknik penyeimbangan data seperti SMOTE atau *class*

weighting dan optimasi parameter secara sistematis. Dengan perbandingan yang terukur, diharapkan dapat ditemukan model dan konfigurasi terbaik dengan performa optimal dari segi akurasi, presisi, *recall*, maupun efisiensi komputasi.

UCAPAN TERIMAKASIH

Selesainya penelitian ini tidak lepas dari bimbingan, dukungan, dan bantuan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan terima kasih ke pada dukungan kedua orang tua selama mengambil data, Ibu Luluk Muthoharoh selaku dosen pembimbing selama penulisan penelitian, Ibu Wiwik Akta Piantri selaku pembimbing lapangan dalam pengambilan data, Pimpinan Kantor Badan Pusat Statistik Republik Indonesia yang telah memberikan kesempatan penulis untuk melakukan magang serta pengambilan data, serta karyawan dan teman-teman saat pengambilan data yang membuat kesempatan ini lebih bermakna.

DAFTAR PUSTAKA

Anggraini, W., Utami, M., Berlianty, J., dkk. (2021). Klasifikasi sentimen masyarakat terhadap kebijakan kartu prakerja di Indonesia. *Faktor Exacta*, *13*(4), 255–261. https://doi.org/10.30998/faktorexacta.v13i4.7964

Arifin, N., Enri, U., Sulistiyowati, N., dkk. (2021). Penerapan algoritma support vector machine (SVM) dengan TF-IDF N-Gram untuk text classification. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, *6*(2). https://doi.org/10.30998/string.v6i2.

Borman, R. I., & Wati, M. (2020). Penerapan data mining dalam klasifikasi data anggota Kopdit Sejahtera Bandarlampung dengan algoritma Naïve Bayes. *Jurnal Ilmiah Fakultas Ilmu Komputer*, *9*(1), 25–34.

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis pembobotan kata pada klasifikasi text mining. *Jurnal Teknologi Informasi*, *3*(2), 179. https://doi.org/10.36294/jurti.v3i2.1077

Fitriyah, N., Warsito, B., & Maruddani, D. A. I. (2020). Analisis sentimen Gojek pada media sosial Twitter dengan klasifikasi support vector machine (SVM). *Jurnal Gaussian*, *9*(3), 376–390. https://doi.org/10.14710/j.gauss.v9i3.28932

Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi berita online dengan menggunakan pembobotan TF-IDF dan cosine similarity. *Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(1), 306–312.

Kalaivani, R., & Marivendan, R. (2021). The effect of stop word removal and stemming in data preprocessing. *Annals of R.S.C.B*, *25*(6), 739–746.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(42). https://doi.org/10.1186/s40537-018-0151-6.

Minatika, I., & Arifudin, R. (2025). Implementation of the term frequency-inverse document frequency method for mental health classification using support vector machine algorithm. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *3*(2). https://doi.org/10.15294/rji.v3i2.1921.

Nero, M. D., & He, J. (2018). Is it necessary: *Quality control* in cataloging? *International Journal of Librarianship*, *3*(2), 85–95. https://doi.org/10.23974/ijol.2018.vol3.2.96

Occhipinti, A., Rogers, L., & Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications*, 201, 117193. https://doi.org/10.1016/j.eswa.2022.117193.

Psaltakis, G., Rogdakis, K., Loizos, M., dkk. (2024). One-vs-one, *One-vs-Rest*, and a novel outcome-driven one-vs-one binary classifiers enabled by optoelectronic memristors towards overcoming hardware limitations in multiclass classification. *Discover Materials*, *4*(7). https://doi.org/10.1007/s43939-024-00077-7.

Putra, J. A., Dharmawan, A., & Gondohanindijo, J. (2024). Sentimen analisis aplikasi Digitalent Mobile menggunakan Naïve Bayes dan SVM dengan ekstraksi fitur TF-IDF. *INTECOMS: Journal of Information Technology and Computer Science*, *7*(4). https://doi.org/10.31539/intecoms.v7i4.11110.

Riadi, I., Umar, R., & Aini, F. D. (2019). Analisis perbandingan detection traffic anomaly dengan metode naive bayes dan support vector machine (SVM). *ILKOM Jurnal Ilmiah*, *11*(1), 17–24. https://doi.org/10.33096/ilkom.v11i1.361.17-24.

Rindiyani, R., Primadewi, A., Maimunah, M., dkk. (2022). Klasifikasi penjualan berdasarkan platform pada UMKM Omah Branded menggunakan random forest. *JURIKOM (Jurnal Riset Komputer)*, *9*(5), 1520. https://doi.org/10.30865/jurikom.v9i5.4949.

Riza, F. (2022). Analisis dan prediksi data penjualan menggunakan machine learning dengan pendekatan ilmu data. *Data Science Indonesia*, *1*(2), 62–68. https://doi.org/10.47709/dsi.v1i2.1308.

Safitri, T., Umaidah, Y., & Maulana, I. (2023). Analisis sentimen pengguna Twitter terhadap grup musik BTS menggunakan algoritma support vector machine. *Journal of Applied Informatics and Computing*, *7*(1), 28–35. https://doi.org/10.30871/jaic.v7i1.5039.

Salam, R. R., Jamil, M. F., Ibrahim, Y., dkk. (2023). Sentiment analysis of cash direct assistance distribution for fuel oil using support vector machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*, 27–35.

Sumitro, P. A., Mulyana, D. I., Saputro, W., dkk. (n.d.). Analisis sentimen terhadap vaksin Covid-19 di Indonesia pada Twitter menggunakan metode lexicon based. *Jurnal Informatika Cipta Karya*.

Sun, A., Lim, E.-P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Information Processing & Management, *45*(5), 559–578. https://doi.org/10.1016/j.ipm.2009.05.001.

Suryati, E., & Aldino, A. A. (2023). Analisis sentimen transportasi online menggunakan ekstraksi fitur model Word2Vec text embedding dan algoritma support vector machine (SVM). *Jurnal Teknologi dan Sistem Informasi (JTSI)*, *4*(1), 96–106. https://doi.org/10.33365/jtsi.v4i1.2445.

Utami, V., & Sihite, D. (2021). Quality management in libraries case study: Book collection retrieval in academic libraries. *Khizanah al-Hikmah: Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan*, *9*, 182. https://doi.org/10.24252/kah.v9cf1

Wiguna, Arya dkk. (2025). Analisis klasifikasi sentimen pengguna media sosial terhadap tagar #IndonesiaGelap pada platform X dengan perbandingan metode klasifikasi SVM, Random Forest, dan Naïve Bayes Classifier. ResearchGate. https://www.researchgate.net/publication/392571720.

Yao, L., Pengzhou, Z., & Chi, Z. (2019). Research on news keyword extraction technology based on TF-IDF and text rank. *In Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science* (ICIS) (pp. 452–455). IEEE. https://doi.org/10.1109/ICIS46139.2019.8940293