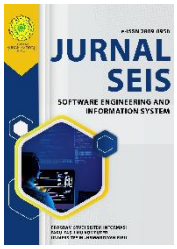
Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index



OPTIMASI KNN DENGAN PSO UNTUK KLASIFIKASI KASUS HUKUM DI AUSTRALIA MENGGUNAKAN N-GRAM

Karan[1*], M. Alidin[2], Rafi Fadilah[3]
[1,2,3]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
email: 220401206@student.umri.ac.id*
email: 220401285@student.umri.ac.id
email: 220401249@student.umri.ac.id

Abstract

This study aims to improve the accuracy of legal case classification in Australia by integrating the K-Nearest Neighbors (KNN) algorithm optimized using Particle Swarm Optimization (PSO) and N-Gram-based text representation. The dataset consists of 15,263 legal documents collected from the Federal Court of Australia (FCA) with an 80:20 data split for training and testing. The classification process is carried out by applying TF-IDF weighting and a combination of N-Gram (unigrams, bigrams, trigrams) to enrich the data representation. The PSO optimization results show an optimal K value of 9, with a testing accuracy reaching 96%. The evaluation of the model performance shows a precision value of 0.95, a recall of 0.96, and an F1-Score of 0.94. These results indicate that the combination of KNN, PSO, and N-Gram is able to significantly improve the performance of legal document classification, especially in the Cited case category. However, the weakness of the model in the Not Cited category indicates the need to develop a method to handle data imbalance in order to improve model generalization.

Keywords: *knn, pso, n-gram, tf-idf, legal classification*

Abstrak

Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi kasus hukum di Australia dengan mengintegrasikan algoritma K-Nearest Neighbors (KNN) yang dioptimalkan menggunakan Particle Swarm Optimization (PSO) dan representasi teks berbasis N-Gram. Dataset terdiri dari 15.263 dokumen hukum yang dikumpulkan dari Pengadilan Federal Australia (FCA) dengan pembagian data 80:20 untuk pelatihan dan pengujian. Proses klasifikasi dilakukan dengan menerapkan pembobotan TF-IDF dan kombinasi N-Gram (unigram, bigram, trigram) untuk memperkaya representasi data. Hasil optimasi PSO menunjukkan nilai K optimal sebesar 9, dengan akurasi pengujian mencapai 96%. Evaluasi kinerja model menunjukkan nilai precision sebesar 0,95, recall sebesar 0,96, dan F1-Score sebesar 0,94. Hasil ini menunjukkan bahwa kombinasi KNN, PSO, dan N-Gram mampu meningkatkan performa klasifikasi dokumen hukum secara signifikan, khususnya pada kategori kasus Cited. Namun, kelemahan model pada kategori Not Cited mengindikasikan perlunya pengembangan metode untuk menangani ketidakseimbangan data guna meningkatkan generalisasi model.

Keywords: knn, pso, n-gram, tf-idf, klasifikasi hukum

PENDAHULUAN

Dengan terus meningkatnya volume data hukum digital di pengadilan federal Australia, proses analisis dan klasifikasi dokumen menjadi semakin kompleks. Ketidakakuratan dalam klasifikasi dapat memperpanjang waktu pengambilan keputusan dan berpotensi mengakibatkan ketidakadilan, sehingga menimbulkan dampak negatif pada sistem hukum dan masyarakat. Dalam konteks ini, pengelompokan dan klasifikasi kasus hukum yang akurat sangat penting untuk membantu hakim dan pengacara menemukan preseden yang relevan secara efisien,

hal ini memungkinkan pengambilan keputusan yang lebih responsif dan presisi [1].

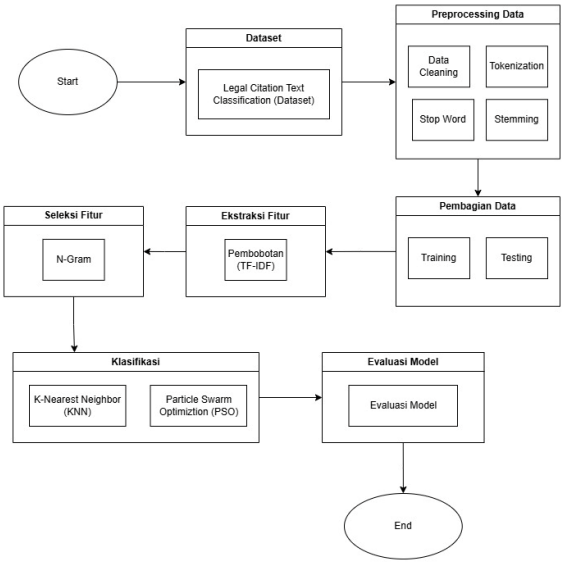
Berbagai metode klasifikasi telah dikembangkan untuk menangani data teks hukum, seperti Support Vector Machine (SVM) [2], Decision Tree [3], dan K-Nearest Neighbors (KNN) [4]. Metode-metode ini telah banyak digunakan, namun memiliki keterbatasan. SVM, meskipun akurat, sering kali memerlukan waktu pelatihan yang panjang untuk dataset berskala besar. Decision Tree terkadang kurang efektif dalam menangani data yang memiliki banyak fitur kontinu. Sementara itu, KNN, yang sederhana dan efektif dalam menangani pola data yang kompleks, memiliki kelemahan dalam hal sensitivitas terhadap parameter seperti nilai K, serta sulit untuk diterapkan pada dataset yang sangat besar [5].

Untuk mengatasi tantangan ini, teknik optimisasi berbasis populasi, seperti Particle Swarm Optimization (PSO), telah terbukti efektif. PSO mampu menghindari jebakan lokal optimum dan menentukan parameter optimal pada algoritma pembelajaran mesin, termasuk KNN [6]. Dibandingkan metode optimasi lainnya seperti Grid Search, yang bersifat eksplisit dan cenderung membutuhkan waktu lebih lama karena melakukan pencarian menyeluruh di seluruh kombinasi parameter, PSO lebih efisien karena mampu memanfaatkan pencarian heuristik berbasis populasi yang adaptif. Selain itu, dibandingkan dengan Genetic Algorithm (GA) yang juga berbasis populasi, PSO memiliki proses yang lebih sederhana karena tidak melibatkan operasi kompleks seperti crossover dan mutation, sehingga lebih sesuai untuk optimasi cepat pada kasus klasifikasi teks hukum berskala besar. Di sisi lain, model N-Gram memungkinkan representasi teks yang lebih baik dengan mempertimbangkan konteks dan pola bahasa dalam dokumen hukum, yang dapat mendukung klasifikasi yang lebih presisi.

Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi kasus hukum di pengadilan federal Australia dengan mengintegrasikan PSO untuk optimasi parameter KNN dan menggunakan model N-Gram sebagai representasi teks. Kombinasi kedua pendekatan ini diantisipasi mampu menawarkan solusi optimal untuk mengatasi tantangan analisis data hukum yang kompleks dan mendukung efisiensi sistem hukum secara keseluruhan.

METODOLOGI PENELITIAN

Penelitian ini dirancang untuk mengklasifikasikan kasus hukum berdasarkan Cited dan Uncited dengan penerapan algoritma KNN yang telah dioptimalkan menggunakan PSO. Proses penelitian ini disusun secara sistematis agar setiap tahap dapat dijalankan secara optimal, sebagaimana dijelaskan dalam diagram alur yang disajikan pada Gambar 1.



Gambar 1 Alur Penelitian

Sebagai tahap awal, penelitian ini melakukan pengumpulan dataset teks hukum yang diambil dari kasus-kasus Pengadilan Federal Australia (FCA) melalui platform AustLII. Dataset ini mencakup kasus dari tahun 2006 hingga 2009 dengan informasi yang meliputi slogan, kalimat kutipan, slogan kutipan, serta kelas kutipan. Dataset ini dipilih untuk memastikan keberagaman data dan relevansi dengan tujuan penelitian. Setelah pengumpulan data, preprocessing diterapkan untuk menjamin kualitas serta keakuratan data. Langkah-langkah preprocessing meliputi pembersihan data untuk menghilangkan karakter khusus dan redundansi, tokenisasi untuk memisahkan teks menjadi unit kata atau frasa, menghapus kata-kata yang dianggap umum dan kurang bermakna (stop words), serta melakukan stemming guna menyederhanakan kata menjadi bentuk dasarnya. Setelah preprocessing, data diolah melalui ekstraksi fitur dengan menerapkan pendekatan TF-IDF untuk menentukan tingkat kepentingan setiap kata dalam dokumen hukum berdasarkan pola kemunculannya. Model N-Gram (unigram, bigram, dan trigram) juga diterapkan untuk menangkap pola bahasa dan konteks yang ada dalam teks, sehingga memperkaya representasi data. Data yang telah diproses dibagi menjadi dua kelompok, yaitu data pelatihan untuk melatih model dan data pengujian untuk menilai kinerja model.

Pada tahap klasifikasi, digunakan algoritma KNN untuk memprediksi kategori. Untuk meningkatkan akurasi model, parameter K pada KNN dioptimalkan menggunakan Particle Swarm Optimization (PSO). PSO bekerja dengan mensimulasikan proses pencarian parameter optimal

berbasis populasi, di mana fungsi objektifnya adalah memaksimalkan akurasi model pada data pelatihan. Setelah model selesai dilatih, evaluasi dilakukan untuk menguji sejauh mana model berfungsi dengan baik. Proses Evaluasi menggunakan metrik seperti akurasi, precision, recall, dan F1-score untuk mengukur kualitas prediksi model. Validasi tambahan dilakukan dengan mengimplementasikan teknik k-Fold Cross-Validation untuk memastikan hasil yang stabil dan bebas dari bias. Pada akhirnya, model yang dihasilkan dioptimalkan untuk memenuhi tujuan penelitian ini, yaitu meningkatkan akurasi klasifikasi kasus hukum dengan mengintegrasikan pendekatan PSO, N-Gram, dan TF-IDF.

1. Dataset

Penelitian ini memanfaatkan data kasus hukum yang tersedia melalui platform AustLII, yang mencakup keputusan-keputusan yang dikeluarkan oleh Pengadilan Federal Australia (FCA) [7]. Dataset mencakup kasus hukum dari tahun 2006 hingga 2009, dengan informasi seperti slogan, kalimat kutipan, slogan kutipan, serta kelas kutipan. Dataset ini dibagi menjadi dua kategori utama, yaitu *Cited* (termasuk *cited* dan *applied*) dan *Uncited* (termasuk *related* dan *distinguished*), untuk mendukung proses klasifikasi berbasis teks.

2. Preprocessing Data

Preprocessing data adalah langkah-langkah pengolahan data yang diawali dengan pemilihan dan pembersihan data (data cleaning) untuk meningkatkan kualitas serta relevansi data sebelum digunakan dalam analisis lebih lanjut [8]. Yang meliputi (a) pemilihan dan pembersihan data (data cleaning) dengan menjadikan seluruh teks dalam bentuk huruf kecil serta menghilangkan karakter yang tidak diperlukan, seperti angka maupun tanda baca, tanpa mengubah makna teks. Langkah ini bertujuan untuk menyederhanakan representasi data dan mengurangi noise pada dataset; (b) data diproses lebih lanjut melalui tokenisasi**,** yaitu dengan memisahkan teks menjadi kata-kata individual serta menghapus spasi berlebih dan karakter tidak relevan, seperti tanda baca, emoji, dan URL [9]. Tokenisasi mempermudah analisis data berbasis teks dengan memisahkan kata-kata yang nantinya digunakan dalam model klasifikasi; (c) pengahapusan stop words menggunakan pustaka NLTK untuk mengeliminasi kata-kata yang umum seperti "dan," "atau," dan "di," yang tidak memiliki pengaruh besar terhadap hasil analisis; dan (d) *stemming*, yaitu menggunakan algoritma Porter Stemmer untuk mengubah kata ke bentuk dasarnya [10]. *Stemming* dilakukan untuk mengurangi dimensi fitur dan memastikan bahwa kata-kata dengan arti yang sama memiliki representasi yang seragam dalam analisis model

3. Pembagian Data

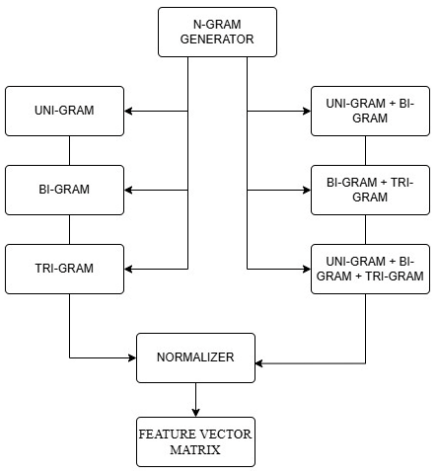
Proses pemisahan dataset terdiri dari dua kelompok utama: data training dan data testing. Data yang dianalisis dalam penelitian ini mencakup teks hukum dengan berbagai atribut yang telah diproses melalui tahap preprocessing sebelumnya. Set data pelatihan digunakan untuk melatih model klasifikasi menggunakan algoritma KNN, sedangkan set data pengujian digunakan untuk mengevaluasi performa model yang telah dilatih. Pemisahan ini bertujuan agar model dapat dilatih secara efektif menggunakan sebagian data, sekaligus memungkinkan evaluasi yang objektif pada data yang belum dikenali oleh model sebelumnya.

4. Ekstraksi Fitur

Setelah melalui tahap preprocessing, teks akan diubah menjadi format numerik melalui ekstraksi fitur agar dapat diproses oleh model klasifikasi. Penelitian ini menggunakan pendekatan *TF-IDF*, yang menilai setiap kata dalam dokumen berdasarkan frekuensi kemunculannya dalam dokumen tertentu dibandingkan dengan keseluruhan koleksi dokumen [11]. Pendekatan ini bertujuan untuk menilai kata-kata dengan relevansi tinggi dalam dokumen tertentu, dan mengurangi kontribusi kata-kata yang umum muncul dalam banyak dokumen.

5. Selektif Fitur

Metode seleksi fitur N-Gram digunakan untuk menangkap pola bahasa dan konteks dalam dokumen hukum dengan mempertimbangkan urutan token (kata atau frasa) di dalam teks. Metode ini memberikan kesempatan untuk analisis lebih mendalam terhadap struktur linguistik teks hukum, yang penting dalam mendukung proses klasifikasi kasus hukum menjadi kategori Cited dan Uncited. Dalam penelitian ini, token yang dianalisis berupa kata-kata, dengan nilai **N** mengacu pada jumlah kata yang dipertimbangkan dalam satu grup (contoh: unigram = satu kata, bigram = dua kata berturutan, trigram = tiga kata berturutan).



Gambar 2 Seleksi Fitur dengan Kombinasi N-Gram

Gambar di atas menjelaskan proses seleksi fitur berbasis N-Gram untuk menghasilkan representasi vektor fitur menggunakan bobot TF-IDF. Dataset yang telah melalui tahap preprocessing diproses lebih lanjut menggunakan fungsi generator TF-IDF Vectorizer dari pustaka Sklearn, yang mendukung kombinasi N-Gram. Penelitian ini menggunakan konfigurasi range N-Gram (1,3) untuk menghasilkan kombinasi Unigram, Bigram, dan Trigram (UniBiTri). Kombinasi ini dipilih karena mampu menangkap informasi dari kata-kata individual sekaligus hubungan antar kata dalam konteks yang lebih luas.

Setiap kombinasi N-Gram kemudian diberikan nilai TF-IDF untuk memberi bobot pada kata atau frasa yang dianggap relevan dalam dokumen tertentu, dan mengurangi bobot kata yang umum muncul di koleksi dokumen lainnya.. Proses ini membantu meningkatkan kualitas fitur dengan memperhatikan relevansi dan signifikansi kata dalam dokumen hukum. Selanjutnya, hasil TF-IDF dari setiap kombinasi N-Gram dinormalisasi untuk membentuk Feature Vector Matrix. Matrix ini merupakan representasi numerik dari teks yang siap digunakan dalam proses klasifikasi. Dalam penelitian ini, jumlah fitur maksimum yang digunakan ditetapkan sebanyak 3000 fitur untuk menjaga efisiensi komputasi sekaligus memastikan kualitas klasifikasi.

6. Klasifikasi

Proses klasifikasi dalam penelitian ini dilakukan menggunakan algoritma KNN yang dioptimalkan dengan PSO. KNN bekerja dengan cara mengklasifikasikan data berdasarkan jarak atau kedekatannya dengan data dalam set pelatihan. Setiap data pada set pelatihan memiliki label kelas, yang digunakan untuk menentukan kategori data uji berdasarkan tetangga terdekatnya. Metode pengukuran jarak yang digunakan meliputi metrik seperti Euclidean atau Manhattan, yang memengaruhi akurasi model KNN. Tahapan klasifikasi melibatkan langkah-langkah berikut:

a. Pengukuran Jarak

Pada langkah awal, algoritma KNN menghitung jarak antara data uji dan seluruh data pelatihan untuk mengidentifikasi tetangga terdekat. Jarak dihitung menggunakan metrik yang dipilih, seperti Euclidean untuk jarak garis lurus atau Manhattan untuk jarak berdasarkan grid. Pemilihan metrik ini berperan penting dalam menentukan akurasi klasifikasi karena hasil klasifikasi sangat bergantung pada kedekatan antar data.

b. Inisialisasi Parameter dengan PSO

Optimasi parameter KNN dimulai dengan Particle Swarm Optimization (PSO), di mana populasi awal partikel diinisialisasi. Setiap partikel mewakili kombinasi parameter KNN, seperti jumlah tetangga (K) dan jenis metrik jarak. Posisi awal partikel diatur secara acak, memberikan variasi konfigurasi parameter yang akan dievaluasi selama proses optimasi.

c. Perbarui Kecepatan dan Posisi Partikel

Selama iterasi, PSO melakukan pembaruan pada kecepatan dan posisi setiap partikel berdasarkan informasi dari konfigurasi terbaik lokal (local best) dan global (global best). Kecepatan partikel menentukan perubahan nilai parameter, sementara posisi partikel menunjukkan parameter KNN yang sedang diuji. Dengan mekanisme ini, PSO secara adaptif mengeksplorasi ruang parameter untuk menemukan konfigurasi optimal.

d. Evaluasi Kinerja Partikel

Setiap kombinasi parameter yang diajukan oleh partikel diuji dengan melatih model KNN menggunakan data pelatihan. Kinerja model dievaluasi berdasarkan metrik tertentu, seperti akurasi pada data validasi. Nilai akurasi ini digunakan untuk menghitung fitness setiap partikel, yang mencerminkan seberapa baik parameter yang diusulkan dalam meningkatkan kinerja model.

e. Optimasi Parameter KNN

PSO terus mengiterasi proses evaluasi dan pembaruan hingga mencapai konfigurasi parameter terbaik. Fokus optimasi adalah menentukan nilai K (jumlah tetangga) dan metrik jarak yang menghasilkan akurasi klasifikasi tertinggi. Dengan demikian, PSO membantu meningkatkan kinerja KNN dengan memastikan bahwa parameter yang dipilih optimal untuk dataset hukum yang digunakan.

7. Evaluasi Model

Pengujian model bertujuan untuk mengevaluasi sejauh mana algoritma KNN dapat membedakan kasus hukum yang dikutip dan tidak dikutip. Evaluasi ini menggunakan beberapa metrik utama, yaitu akurasi, precision, recall, dan F1-score [12]. Akurasi menunjukkan persentase keseluruhan prediksi yang tepat berdasarkan data uji. Sementara itu, precision mengukur sejauh mana model mampu mengidentifikasi data positif dengan benar dibandingkan dengan seluruh prediksi positif yang dibuat. Recall menunjukkan seberapa baik model dalam mengenali semua sampel yang termasuk dalam kategori positif, sementara F1-score merupakan metrik gabungan yang mempertimbangkan precision dan recall, terutama dalam kondisi dataset dengan distribusi kelas yang tidak merata.

Confusion Matrix digunakan untuk memberikan analisis mendalam mengenai hasil klasifikasi [13]. Evaluasi menggunakan Confusion

Matrix merupakan suatu metode yang memungkinkan untuk mencatat sejauh mana akurasi atau kesalahan prediksi suatu algoritma dalam proses klasifikasi [14]. Confusion Matrix membantu memvisualisasikan distribusi prediksi model dalam empat kategori, yaitu TP menunjukkan data yang benar-benar positif dan diklasifikasikan dengan tepat, TN merepresentasikan data negatif yang berhasil diidentifikasi sebagai negatif, FP terjadi ketika data yang seharusnya negatif diklasifikasikan sebagai positif dan FN terjadi ketika data positif yang teridentifikasi secara salah sebagai negatif [15].

Table . *Confussion matrix*

Predict	**Actual**	
	True	False
True	True Positif (TP)	False Negatif (FN)
False	False Positif (TF)	True Negatif (TN)

Hasil evaluasi ini memberikan pemahaman menyeluruh terkait kinerja model serta membantu dalam mengidentifikasi pola kesalahan yang mungkin terjadi. Dengan demikian, langkah-langkah perbaikan dapat dilakukan untuk mengoptimalkan kemampuan model dalam proses klasifikasi.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Kumpulan data diperoleh dari Pengadilan Federal Australia (FCA) melalui platform AustLII, yang mencakup kasus-kasus hukum dari tahun 2006 hingga 2009. Dataset ini terdiri dari total 15.263 dokumen hukum yang diklasifikasikan ke dalam dua kategori utama, yaitu *cited* (termasuk *cited* dan *applied*) sebanyak 14.548 dokumen dan uncited (termasuk *related* dan *distinguished*) sebanyak 715 dokumen.

Table 2 Dataset

No	**Case Outcome**	**Case Title**	**Case Text**
1	Cited	Alpine Hardwood (Aust) Pty Ltd v ..	Ordinarily that discretion will ...
2	Cited	Black v Lipovac [1998] FCA 699 ; (1998) 217 ALR..	The general principles governing the ...
...	...	...	...
25205	Cited	Spiel v Commodity Brokers Australia Pty Ltd (In liq)	Once the threshold prescribed by s ..
25206	Distinguished	Tullock Ltd v Walker (Unreported, Su...	Given the extent to which Deumer stands to gain from t...
25207	Distinguished	Yandil Holdings Pty Ltd v Insurance Co of North America ...	In my view, it is clear that the Court may do ..

Setelah data dikumpulkan, tahap preprocessing dilakukan untuk meningkatkan kualitas dan relevansi data. Proses ini melibatkan normalisasi teks, seperti pengubahan menjadi huruf kecil, penghapusan angka, karakter khusus, dan teks redundan. Hasil preprocessing menunjukkan distribusi data yang telah dikelompokkan ke dalam kategori *Cited* dan *Uncited*. Hasil normalisasi data ditampilkan pada Tabel 3.

Tabe 3 Hasil normalisasi

Cited		
No	**Clean Text**	**Label**
1	ordinarili discret exercis cost follow event award parti parti basi departur normal practic award indemn cost requ...	Cited
2	ener principl govern exercis discret award indemn cost reject unsuccess parti call calderbank letter set judgment full ...	Cited
3	ordinarili discret exercis cost follow event award parti parti basi departur normal practic award indemn cost requ...	Cited
4	gener principl govern exercis discret award indemn cost reject unsuccess pa...	Cited
...	...	....
Uncited		
No	**Clean Text**	**Label**
1	june made order extend conven period meet creditor requir held corpor act cth act midnight juli relat number compani...	Uncited
2	june made order extend conven period meet creditor requir held corpor act cth act midnight juli relat number compani ...	Uncited
3	june made order extend conven period meet creditor requir held corpor act cth act midnight juli relat number compani ...	Uncited
4	applic submit question whether treasur satisfi acquisit contrari nation interest j...	Uncited
...	...	...

Data kemudian dipisahkan menjadi dua kelompok, yaitu data pelatihan, yang digunakan untuk membangun model dan data uji digunakan mengukur performanya. Dataset terdiri dari beberapa parameter yang telah ditentukan sebelumnya, di mana teks hasil preprocessing digunakan sebagai variabel x, sementara label kutipan (*Cited* dan *Uncited*) berfungsi sebagai variabel y.

2. Pembobotan TF-IDF dengan N-gram

Proses ini menghasilkan bobot yang mencerminkan relevansi setiap term dalam sebuah dokumen relatif terhadap seluruh kumpulan dokumen. Dalam penelitian ini, pendekatan N-Gram digunakan untuk menghasilkan representasi term yang lebih kaya. Kombinasi unigram, bigram, dan trigram diterapkan untuk menangkap pola-pola penting dalam teks hukum, seperti istilah spesifik dan kombinasi kata yang relevan. Hasil dari perhitungan ini memberikan gambaran tentang kontribusi setiap term dalam analisis teks dan disajikan pada Tabel 4.

Table 4 Nilai TF-idf f pada Data Training dan Testing dengan N-gram

	TF-IDF							
Term	**Data Training**					**Data Testing**		
	D1	**D2**	**D3**	**D4**	**D5**	**U1**	**U2**	**U3**
action	0,058885	0	0	0	0	0	0	0
action against	0,058885	0	0	0	0	0	0	0
after	0,015836	0	0	0.040186	0	0,040186	0	0
against	0,058885	0	0	0	0	0	0	0
adverse	0,058885	0	0	0	0	0	0	0

Berdasarkan Tabel di atas, hasil perhitungan nilai TF-IDF menunjukkan distribusi bobot antar dokumen pada data pelatihan dan pengujian dengan pendekatan N-Gram. Pada kolom Data Training, term "after" memiliki bobot sebesar 0,040186 pada dokumen D4, sementara dokumen lainnya seperti D1, D2, D3, dan D5 sebagian besar tidak memiliki bobot signifikan untuk term tersebut. Dokumen D1, di sisi lain, memiliki beberapa term dengan bobot yang relatif konsisten, seperti "action" sebesar 0,058885, "action against" sebesar 0,058885, dan "against" sebesar 0,058885. Pada kolom Data Testing, term "after" juga memiliki bobot sebesar 0,040186 pada dokumen U1, sementara term lainnya seperti "action", "action against", dan "against" tidak menunjukkan bobot yang signifikan pada data pengujian. Hasil ini menunjukkan bahwa penggunaan N-Gram (unigram, bigram, dan trigram) memberikan kontribusi yang berbeda-beda dalam merepresentasikan dokumen tertentu, mencerminkan pentingnya pola-pola linguistik dalam analisis teks hukum.

3. Klasifikasi

Penelitian ini menerapkan algoritma KNN dalam proses klasifikasi, yang selanjutnya dioptimalkan dengan PSO guna meningkatkan akurasi prediksi. Data teks hukum yang telah melalui proses pembobotan TF-IDF dengan kombinasi N-Gram (unigram, bigram, trigram) diproses lebih lanjut sebagai masukan bagi model klasifikasi. Evaluasi kinerja model dilakukan dengan membandingkan tiga rasio pembagian data, yaitu 80:20, 70:30, dan 60:40, untuk menilai pengaruh proporsi data latih terhadap performa model. Penggunaan N-Gram memungkinkan model untuk menangkap pola kata individual dan kombinasi frasa yang lebih kompleks, seperti istilah hukum spesifik yang muncul dalam dokumen. Selain itu, optimasi parameter model, seperti jumlah tetangga (K) dan metrik jarak (Euclidean), menggunakan PSO membantu menentukan konfigurasi optimal untuk meningkatkan akurasi klasifikasi. Hasil akurasi dari berbagai konfigurasi N-Gram dan pembagian data ditampilkan pada Tabel 5 berikut.

Table 5 Hasil Akurasi Model KNN dengan Optimasi PS

Pembagian Data	**Unigram**	**Bigram**	**Trigram**	**Unigram+ Bigram+Trigram**
80:20	0.471683	0.475886	0.478887	0.475485
70:30	0.467716	0.489061	0.494664	0.471051
60:40	0.453772	0.467881	0.470382	0.455874

Berdasarkan hasil pada Tabel 5, terlihat bahwa trigram memberikan akurasi tertinggi pada proporsi data 70:30 dengan nilai 0.494664, diikuti oleh bigram dengan nilai 0.489061. Kombinasi N-Gram (unigram, bigram, trigram) menunjukkan performa yang cukup kompetitif, meskipun sedikit lebih rendah dibandingkan trigram pada proporsi data yang sama. Pada proporsi data 80:20, trigram tetap unggul dengan akurasi sebesar 0.478887, sementara pada proporsi data 60:40, performa model cenderung menurun pada semua konfigurasi, mengindikasikan bahwa ukuran data pelatihan yang lebih kecil memengaruhi efektivitas model dalam mengenali pola.

Penurunan akurasi pada proporsi data 60:40 kemungkinan disebabkan oleh berkurangnya representasi pola-pola linguistik dalam data pelatihan. Hal ini terutama terlihat pada kombinasi N-Gram, yang cenderung membutuhkan data pelatihan lebih banyak untuk menangkap pola yang kompleks. Sebaliknya, trigram tetap menunjukkan konsistensi karena mampu menangkap frasa spesifik yang lebih relevan dalam dokumen hukum, meskipun tetap dipengaruhi oleh jumlah data yang terbatas.

Penggunaan PSO dalam optimasi parameter KNN terbukti efektif dalam meningkatkan akurasi klasifikasi. Dengan mengoptimalkan jumlah

tetangga (K) dan metrik jarak, PSO membantu model menemukan konfigurasi parameter terbaik yang dapat memaksimalkan akurasi. Namun, hasil ini juga menunjukkan bahwa kombinasi N-Gram dengan model KNN memiliki batasan ketika data pelatihan terlalu kecil, yang menekankan pentingnya proporsi data latih yang mencukupi untuk mencapai hasil optimal.

Evaluasi kinerja model, yang dilakukan dengan memanfaatkan metrik precision, recall, dan F1-score untuk mendapatkan pemahaman yang lebih mendalam tentang sejauh mana kemampuan model dalam mengklasifikasikan dokumen hukum. Selain itu, analisis confusion matrix akan disajikan untuk menggambarkan prediksi benar dan salah pada kategori Cited dan Not Cited, sehingga memberikan wawasan tambahan tentang distribusi kesalahan prediksi model.

4. Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan untuk menilai efektivitas algoritma KNN yang dioptimalkan menggunakan PSO. Data yang digunakan dalam penelitian ini dibagi 80:20 antara pengujian dan pelatihan. Proses optimasi PSO menemukan nilai K optimal sebesar K=9, dengan tingkat akurasi pada data pengujian sebesar 96%.

Table 6. Hasil Evaluasi Metrik untuk Model KNN dengan Optimasi PSO

Kategori	Precision	Recall	F1-Score
Cited	0,96	1,00	0,98
Not Cited	0,71	0,04	0,08
Rata-Rata	0,95	0,96	0,94

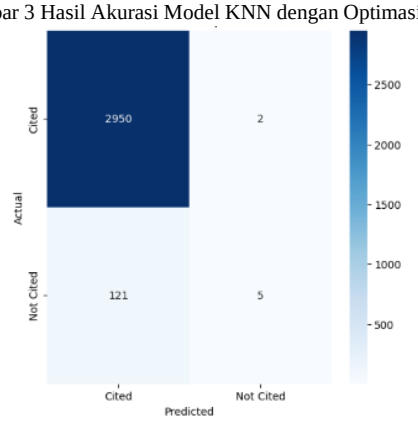
Hasil evaluasi menunjukkan bahwa kategori Cited menunjukkan performa luar biasa, dengan nilai Precision 0.96, Recall 1.00, dan F1-Score 0.98. Namun, pada kategori Not Cited memiliki performa yang lebih rendah dengan nilai Precision 0.71, Recall 0.04, dan F1-Score 0.08. Secara keseluruhan, model memiliki rata-rata Precision 0.95, Recall 0.96, dan F1-Score 0.94.

Berdasarkan hasil ini, algoritma KNN yang dioptimalkan menggunakan PSO mampu menangani pembagian data 80:20 dengan performa yang sangat baik pada kategori Cited. Namun, performa yang sangat rendah pada kategori Not Cited menunjukkan adanya indikasi bahwa model kesulitan membedakan dokumen dalam kategori ini.

Salah satu faktor yang berkontribusi pada rendahnya Recall di kategori Not Cited adalah karakteristik teks dalam dokumen tersebut. Berdasarkan analisis data, dokumen Not Cited cenderung memiliki panjang yang lebih pendek, penggunaan terminologi hukum yang lebih umum, serta pola kalimat yang lebih bervariasi dibandingkan dokumen Cited yang lebih konsisten menggunakan istilah-istilah spesifik yang relevan dengan kasus rujukan. Selain itu, proporsi dokumen Not Cited yang jauh lebih kecil dibandingkan Cited menyebabkan model cenderung menggeneralisasi pola dari kategori mayoritas, sehingga meningkatkan risiko misclassifications pada kategori minoritas.

Ketidakseimbangan data (class imbalance) juga menjadi faktor utama yang memengaruhi rendahnya performa recall. Seperti dijelaskan oleh Moeslem et al (2025), model pembelajaran mesin sering menunjukkan bias ke kelas mayoritas dalam kondisi distribusi data yang tidak seimbang, sehingga prediksi terhadap kelas minoritas menjadi tidak akurat [16]. Permasalahan ini perlu menjadi perhatian dalam pengembangan model di masa depan agar model memiliki kemampuan generalisasi yang lebih baik.

Gambar 3 Hasil Akurasi Model KNN dengan Optimasi PS



Berdasarkan hasil Confusion Matrix, model terbukti sangat efektif dalam mengidentifikasi kategori Cited, dengan 2950 dokumen dari kategori ini berhasil diprediksi dengan benar sebagai Cited (True Positive), sementara hanya 2 dokumen yang salah diprediksi sebagai Not Cited (False Negative). Namun, model memiliki kelemahan dalam mengenali kategori Not Cited, di mana hanya 5 dokumen yang berhasil diprediksi dengan benar (True Negative), sementara 121 dokumen keliru diprediksi sebagai Cited (False Positive). Kesalahan ini menunjukkan bahwa model cenderung over-predict pada kategori Cited, yang kemungkinan disebabkan oleh ketidakseimbangan jumlah data antara kategori Cited dan Not Cited. Secara keseluruhan, model bekerja sangat baik untuk kategori Cited tetapi memerlukan peningkatan dalam menangani kategori Not Cited.

SIMPULAN DAN SARAN

Penelitian ini telah berhasil menerapkan algoritma KNN yang dioptimalkan menggunakan PSO untuk klasifikasi kasus hukum di Australia dengan menggunakan representasi teks berbasis N-Gram (unigram, bigram, trigram, dan kombinasi). Dengan pembagian data 80:20 untuk pelatihan dan pengujian, nilai optimal K=9 yang ditemukan menggunakan PSO menghasilkan akurasi pengujian

sebesar 96%. Berdasarkan evaluasi metrik, model menunjukkan performa yang sangat baik dalam kategori Cited, dengan nilai Precision 0.96, Recall 1.00, dan F1-Score mencapai 0.98. Namun, kelemahan model terlihat pada kategori Not Cited, dengan nilai Recall 0.04 dan F1-Score 0.08, yang disebabkan oleh ketidakseimbangan data antara kedua kategori. Hasil ini menunjukkan bahwa optimasi PSO dapat meningkatkan performa model pada kategori dominan, namun memerlukan strategi tambahan untuk menangani ketidakseimbangan data.

Untuk meningkatkan kinerja model dalam penelitian selanjutnya, disarankan untuk menggunakan teknik penyeimbangan data seperti oversampling pada kategori Not Cited atau undersampling pada kategori Cited guna mengurangi ketidakseimbangan yang memengaruhi performa model. Selain itu, pendekatan cost-sensitive learning juga dapat dipertimbangkan, yaitu dengan memberikan bobot kesalahan (misclassification cost) yang lebih tinggi pada kategori Not Cited, sehingga model lebih memperhatikan dan meningkatkan sensitivitas terhadap kategori minoritas tersebut, meskipun jumlahnya kecil. Selain strategi balancing dan cost-sensitive learning, eksplorasi fitur tambahan seperti metadata hukum, penggunaan algoritma hybrid yang menggabungkan KNN dengan model lain, serta pengujian pada dataset yang lebih luas, diharapkan dapat membantu meningkatkan akurasi dan generalisasi model. Pendekatan lain yang dapat dilakukan adalah penerapan teknik Explainable AI (XAI) untuk memahami lebih dalam bagaimana model mengklasifikasikan dokumen hukum, mengidentifikasi fitur-fitur yang paling berpengaruh, serta menemukan area yang memerlukan peningkatan di masa mendatang.

TERIMA KASIH

Penulis menyampaikan rasa terima kasih kepada Universitas Muhammadiyah yang telah memberikan dukungan dan bimbingan yang sangat penting dalam penyusunan artikel ini. Ucapan terima kasih khusus juga disampaikan kepada Kaprodi Teknik Informatika, Yulia Fatma, S. Kom, M. Cs, atas arahan dan motivasi yang telah diberikan. Bantuan dari para dosen pembimbing, staf pengajar, dan rekan-rekan di lingkungan universitas telah memberikan kontribusi yang signifikan dalam menyelesaikan penelitian ini. Penulis berharap artikel ini dapat membantu memajukan ilmu pengetahuan dengan cara yang konstruktif, terutama dalam bidang klasifikasi teks hukum.

DAFTAR PUSTAKA

[1] A. K. Jailani Tanjung, H. Purwadi, and , Hartiwiningsih, “Paradigma Hakim Dalam Memutuskan Perkara Pidana Di Indonesia,” *J. Huk. dan Pembang. Ekon.*, vol. 7, no. 1, p. 39, 2019, doi: 10.20961/hpe.v7i1.29178.

[2] G. Dlamini, Z. Kholmatova, A. Kruglov, G. Succi, H. Tarasau, and A. Valeev, “Meta-analytical Comparison Of SVM and KNN for Text Classification,” *Int. Conf. "Nonlinearity, Inf. Robot. NIR*, 2021, doi: 10.1109/NIR52917.2021.9666133.

[3] J. González-González, F. de Arriba-Pérez, S. García-Méndez, A. Busto-Castiñeira, and F. J. González-Castaño, “Automatic explanation of the classification of Spanish legal judgments in jurisdiction-dependent law categories with tree estimators,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 7, 2023, doi: 10.1016/j.jksuci.2023.101634.

[4] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications,” *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00973-y.

[5] A. K. Iman, E. Iman, H. Ujianto, F. Sains, and U. T. Yogyakarta, “Analisis Sentimen Pemindahan Ibu Kota Indonesia Menggunakan K-Nearest Neighbor Sentiment Analysis of the Relocation of Indonesia ’ s Capital Using K-Nearest Neighbor,” *J. Pendidik. dan Teknol. Indones.*, vol. 4, no. 12, pp. 759–768, 2024.

[6] M. Rizki, A. Hermawan, and D. Avianto, “Optimization of Hyperparameter K in K-Nearest Neighbor Using Particle Swarm Optimization,” *JUITA J. Inform.*, vol. 12, no. 1, p. 71, 2024, doi: 10.30595/juita.v12i1.20688.

[7] S. Bansal, “Klasifikasi Teks Kutipan Hukum,” p. 2021, 2021, [Online]. Available: https://www.kaggle.com/datasets/shivamb/legal-citation-text-classification/data

[8] M. Muharrom, “Komparasi Algoritma Klasifikasi Naive Bayes Dan K-Nearest Neighbors Dalam Analisis Sentimen Terhadap Opini Film Pada Twitter,” *J. Inform. Dan Tekonologi Komput.*, vol. 3, no. 1, pp. 43–50, 2023, doi: 10.55606/jitek.v3i1.1147.

[9] A. Firdaus, “Aplikasi Algoritma K-Nearest Neighbor pada Analisis Sentimen Omicron Covid-19,” *J. Ris. Stat.*, pp. 85–92, 2022, doi: 10.29313/jrs.v2i2.1148.

[10] G. Putra, A. Brahmantha, E. Utami, and A. Yaqin, “Klasifikasi Genre Anime Berdasarkan Sinopsis Menggunakan Algoritma K-Nearest Neighbors,” *J. Manaj. Inform. Sist. Inf.* , vol. 7, no. 1, pp. 15–24, 2024.

[11] S. N. Yanti, Yuhandri, and Sumijan, “Jurnal KomtekInfo Implementasi K-Nearest Neighbor Berbasis Particle Swarm,” *J. KomtekInfo*, vol. 11, no. 4, pp. 371–379, 2024, doi: 10.35134/komtekinfo.v11i4.586.

[12] M. A. Satriawan and W. Widhiarso, "Klasifikasi Pengenalan Wajah Untuk Mengetahui Jenis Kelamin Menggunakan Metode Convolutional Neural Network," *J. Algoritm.*, vol. 4, no. 1, pp. 43–52, 2023, doi: 10.35957/algoritme.xxxx.

[13] R. Riskawati, F. Fatihanursari, I. Iin, and A. Rizki Rinaldi, "Penerapan Metode Naïve Bayes Classifier Pada Analisis Sentimen Aplikasi Gopay," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 346–353, 2024, doi: 10.36040/jati.v8i1.8699.

[14] S. Agustian *et al.*, "Jurnal Computer Science and Information Technology (CoSciTech) Pengaruh Agregasi Data pada Klasifikasi Sentimen untuk Dataset Terbatas Menggunakan SGD Effect of Data Aggregation on Sentiment Classification for Limited Datasets Using SGD Classifier," *J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 626–634, 2024.

[15] N. Knn, "Analisa Sentimen Pada Media Sosial ' X ' Pencarian Keyword ChatGPT Menggunakan Algoritma K-Nearest Abstrak," vol. 5, no. 3, pp. 3291–3305, 2024.

[16] I. Moeslem *et al.*, "Deep Learning Dengan Teknik Early Stopping Untuk Mendeteksi Deep Learning With Early Stopping Technique For Malware Detection On Iot Devices," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 1, pp. 21–30, 2025, doi: 10.25126/jtiik.2025128267.