Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 13, No. 3, Juni 2026, hlm. 571-580 e-ISSN: 2528-6579

PENINGKATAN SENSITIVITAS SUPPORT VECTOR MACHINE PADA KLASIFIKASI KALIMAT BAKU DAN TIDAK BAKU BAHASA INDONESIA

Rianto Rianto*[1], Eko Setyo Humanika[2], Iwan Hartadi Tri Untoro[3]

[1,2,3] Universitas Teknologi Yogyakarta, Yogyakarta
Email: [1]rianto@staff.uty.ac.id, [2]eko.humanika@staff.uty.ac.id, [3]iwan.hartadi@staff.uty.ac.id
*Penulis Korespondensi

(Naskah masuk: 24 Juni 2025, diterima untuk diterbitkan: 30 November 2025)

Abstrak

Kalimat dalam bahasa Indonesia dapat diklasifikasikan menjadi dua jenis, yaitu kalimat baku dan tidak baku. Kalimat baku digunakan sebagai bahasa resmi dalam acara formal, sementara kalimat tidak baku umum ditemukan dalam komunikasi sehari-hari. Perkembangan teknologi digital turut mendorong pergeseran penggunaan bahasa, sehingga banyak kalimat tidak baku muncul dalam konteks formal. Penelitian ini bertujuan untuk meningkatkan akurasi model klasifikasi kalimat baku dan tidak baku berbahasa Indonesia. Metode yang digunakan adalah *supervised learning* dengan *Support Vector Machine* (SVM) menggunakan *dataset* berjumlah 2.000 kalimat. Peningkatan akurasi dilakukan melalui TF-IDF *Vectorizer* dan augmentasi data. Hasil penelitian menunjukkan bahwa sebelum improvisasi, akurasi model mencapai 98.6% dengan total empat kesalahan klasifikasi pada kalimat baku pendek dan ambigu. Setelah improvisasi, akurasi meningkat menjadi 99.3%, dengan jumlah kesalahan total menurun menjadi tiga. Kebaruan penelitian ini terletak pada fokus klasifikasi kalimat baku dan tidak baku dalam konteks bahasa Indonesia, yang masih jarang dieksplorasi. Kontribusinya adalah menyediakan model yang dapat menjadi dasar aplikasi pemeriksa tata bahasa atau sistem penyaringan teks formal. Namun, penelitian ini memiliki keterbatasan pada augmentasi data yang masih dilakukan secara manual. Penelitian selanjutnya diharapkan dapat mengembangkan sistem augmentasi otomatis serta menambah jumlah *dataset* untuk meningkatkan generalisasi model dan penerapan lebih luas dalam NLP Indonesia.

Kata kunci: *augmentasi data, bahasa indonesia, SVM, TF-IDF vectorizer*

ENHANCING SUPPORT VECTOR MACHINE SENSITIVITY FOR STANDARD AND NON-STANDARD SENTENCES CLASSIFICATION IN INDONESIAN

Abstract

Sentences in Indonesian can be classified into two types: standard and non-standard. Standard sentences are used as the official language in formal events, while non-standard sentences are used in daily conversation. The development of digital technology has contributed to the language shift, which causes many non-standard sentences to be used in standard contexts. This study aims to improve the accuracy of the classification model for standard and non-standard sentences in Indonesian. The method used is supervised learning with a Support Vector Machine (SVM) using a dataset of 2,000 sentences consisting of standard and non-standard forms. Accuracy improvement is carried out through TF-IDF Vectorizer and data augmentation. The results showed that the model accuracy reached 98.6% before improvisation, with four misclassifications in short and ambiguous standard sentences. After improvisation, the accuracy increased to 99.3%, with the number of errors decreasing to three. The novelty of this study lies in its focus on addressing sentence classification into standard and non-standard forms in the Indonesian language, which has rarely been explored compared to other NLP tasks. Its practical contribution is to provide a baseline model that can support grammar-checking tools or text filtering systems in digital platforms. However, this study has limitations in data augmentation, which still needs to be done manually. Future research is expected to develop an automatic augmentation system and increase the number of datasets to improve model generalization and broaden its application in Indonesian NLP.

Keywords: *data augmentation, indonesian, SVM, TF-IDF vectorizer*

1. PENDAHULUAN

Klasifikasi teks merupakan salah satu cabang penting dalam pengolahan bahasa alami atau *Natural Language Processing* (NLP). NLP banyak digunakan dalam berbagai aplikasi, seperti analisis sentimen, deteksi spam, atau kategorisasi dokumen (Triloka dkk, 2022). Dalam konteks bahasa Indonesia,

klasifikasi kalimat baku dan tidak baku menjadi tantangan tersendiri karena karakteristik unik bahasa tersebut, seperti penggunaan bahasa sehari-hari yang sering bercampur dengan kosakata baku (Nurliza dkk, 2025) (Khairul & Perdana, 2024). Klasifikasi ini penting, terutama dalam aplikasi seperti sistem *review* produk dengan analisis sentimen (Liu dkk, 2022).

Namun, tantangan utama dalam klasifikasi ini adalah bagaimana model pembelajaran mesin dapat membedakan antara kalimat baku dan tidak baku, terutama ketika kedua jenis kalimat tersebut memiliki pola yang mirip. Sebagai contoh, kalimat baku yang pendek atau memiliki gaya puitis sering kali sulit dibedakan dari kalimat tidak baku. Masalah ini dapat mengakibatkan kesalahan klasifikasi yang signifikan, terutama dalam aplikasi yang membutuhkan tingkat akurasi tinggi (Rianto dkk, 2021)

Salah satu metode yang sering digunakan untuk tugas klasifikasi adalah *Support Vector Machine* (SVM), yang dikenal karena kemampuannya dalam menangani data set dengan dimensi tinggi. Namun, performa SVM sangat bergantung pada representasi data yang digunakan. Representasi sederhana seperti *CountVectorizer* sering kali tidak cukup untuk menangkap kompleksitas kalimat baku, sehingga memerlukan pendekatan yang lebih canggih (Goyal, 2021).

Penelitian ini mengusulkan penggantian *CountVectorizer* dengan TF-IDF *Vectorizer* untuk meningkatkan sensitivitas model terhadap kata-kata unik dan informatif. Selain itu, augmentasi data dengan menambahkan kalimat baku serupa yang mencerminkan pola ambigu, seperti kutipan atau ungkapan puitis, diharapkan dapat memperluas cakupan pola pelatihan model. Dengan pendekatan ini, diharapkan model dapat mengatasi kesalahan klasifikasi pada kalimat baku dan meningkatkan akurasi keseluruhan dalam klasifikasi kalimat baku dan tidak baku berbahasa Indonesia (Danyal dkk, 2024).

Pada dasarnya klasifikasi kalimat baku dan tidak baku memiliki aplikasi luas dalam berbagai sektor. Sebagai contoh dalam analisis dokumen resmi, klasifikasi ini dapat membantu menyaring dokumen yang menggunakan gaya bahasa baku dengan tepat. Selain itu, untuk layanan pelanggan berbasis *chatbot*, kemampuan mendeteksi dan merespons kalimat baku dan tidak baku dengan tepat dapat meningkatkan kualitas interaksi antara pengguna dan sistem (Rianto dkk, 2020). Pada sektor pendidikan, kemampuan ini juga dibutuhkan untuk mengevaluasi dokumen akademik yang harus menggunakan bahasa baku.

Meskipun berbagai pendekatan telah dilakukan untuk menangani klasifikasi teks, sebagian besar penelitian masih fokus pada bahasa Inggris atau bahasa lain yang memiliki struktur berbeda dengan bahasa Indonesia. Hal ini menciptakan kesenjangan penelitian yang perlu diisi, terutama dalam menangani karakteristik unik bahasa Indonesia. Penelitian ini tidak hanya berkontribusi pada pengembangan metodologi klasifikasi teks dalam bahasa Indonesia tetapi juga menawarkan pendekatan praktis dengan mempertimbangkan keterbatasan sumber daya komputasi yang sering kali menjadi kendala di berbagai lingkungan aplikasi.

Dengan pendekatan yang diusulkan, penelitian ini dapat memberikan kontribusi signifikan dalam pengembangan teknologi pengolahan bahasa alami untuk bahasa Indonesia, khususnya dalam klasifikasi teks. Pendekatan berbasis SVM yang dikombinasikan dengan representasi data yang lebih baik dan augmentasi manual menjadi solusi yang dapat diimplementasikan secara praktis, serta dapat dikembangkan lebih lanjut melalui teknologi augmentasi otomatis.

Penelitian ini dirancang sebagai upaya awal atau *proof of concept* untuk menguji efektivitas pendekatan berbasis Support Vector Machine (SVM) dalam klasifikasi kalimat baku dan tidak baku berbahasa Indonesia. Dengan menggunakan data set berukuran kecil hingga menengah yang terdiri dari 2.000 kalimat, penelitian ini mengeksplorasi kombinasi representasi data berbasis TF-IDF *Vectorizer* dan augmentasi data manual untuk meningkatkan sensitivitas model terhadap pola teks yang kompleks.

Pendekatan ini dipilih dengan mempertimbangkan efisiensi komputasi dan kesesuaian dengan skenario terbatas di mana sumber daya yang tersedia sering kali tidak mencukupi untuk melatih model modern seperti *Transformer*. Dengan demikian, penelitian ini memberikan alternatif yang efisien dan terukur untuk skenario dengan keterbatasan data dan infrastruktur, sekaligus membuka peluang untuk pengembangan lebih lanjut melalui pendekatan yang lebih kompleks.

2. TINJAUAN PUSTAKA DAN TEORI

Pengembangan penelitian memerlukan landasan teori dan pemahaman terhadap penelitian-penelitian terdahulu. Tinjauan pustaka dapat membantu mengidentifikasi kesenjangan penelitian yang ada, sementara kerangka teori menjadi dasar bagi pendekatan yang digunakan. Bab ini memaparkan penelitian-penelitian relevan dalam klasifikasi kalimat baku dan tidak baku, serta menyajikan teori-teori yang mendukung strategi perbaikan melalui TF-IDF dan augmentasi data.

2.1. Tinjauan Pustaka

Klasifikasi teks telah menjadi salah satu bidang utama dalam pengolahan bahasa alami (NLP), dengan penelitian yang beragam baik pada bahasa Inggris

maupun bahasa non-Inggris. *Support Vector Machine* (SVM) telah menunjukkan keberhasilan dalam berbagai tugas klasifikasi teks (Hana dkk, 2020). SVM digunakan dengan *CountVectorizer* untuk klasifikasi analisis sentimen. Namun, metode ini menunjukkan keterbatasan dalam menangani pola teks yang ambigu. Hasil penelitian lain menemukan bahwa penggunaan TF-IDF dapat meningkatkan sensitivitas model terhadap kata-kata unik yang penting, terutama dalam dokumen formal. Meskipun demikian, sebagian besar studi tersebut difokuskan pada bahasa Inggris, sehingga aplikasinya pada bahasa Indonesia memerlukan validasi lebih lanjut (Bengesi dkk, 2023).

Hasil penelitian lain menyebutkan bahwa augmentasi data juga dapat memberikan kontribusi penting. Augmentasi data berbasis sinonim dapat meningkatkan akurasi klasifikasi teks pendek. Penelitian ini membuka peluang untuk mengeksplorasi teknik serupa dalam konteks bahasa Indonesia, khususnya pada kalimat baku dan tidak baku (Bayer dkk, 2022) (Bayer dkk, 2023). Hasil penelitian lain juga mengindikasikan bahwa model berbasis *embedding* seperti *Word2Vec* yang menawarkan fleksibilitas masih kalah relevan dibandingkan dengan pendekatan berbasis fitur sederhana dengan optimasi tertentu.

Penelitian oleh Gabriela dan kawan-kawan (2021) lebih jauh menekankan pada integrasi TF-IDF dengan fitur sintaksis, yang meningkatkan akurasi tetapi mengorbankan waktu pelatihan. Hal ini relevan dengan penelitian Haitao dan kawan-kawan (2020), yang mencatat bahwa klasifikasi kalimat tidak baku dalam media sosial sering bias terhadap panjang teks. Untuk mengatasi bias ini, Abdollahi dan kawan-kawan (2021) mengusulkan augmentasi berbasis sinonim. Hasil penelitian lain menyarankan pendekatan hibrida antara TF-IDF dan *CountVectorizer* untuk mencapai akurasi lebih tinggi hingga 5% (Bilgen & Kaya, 2024).

Augmentasi data juga sering dikaitkan dengan data yang tidak seimbang. Data tidak seimbang sering menjadi tantangan utama dalam klasifikasi kalimat baku, sehingga diperlukan augmentasi data yang terarah. Selain augmentasi, adaptasi model juga diperlukan untuk menangani keunikan bahasa non-Inggris seperti bahasa Indonesia, termasuk dalam representasi teks yang kompleks (Qiu dkk, 2020).

2.2. Support Vector Machine

Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang sering digunakan untuk tugas klasifikasi dan regresi. SVM bekerja dengan mencari *hyperplane* yang optimal untuk memisahkan data dari kelas yang berbeda. *Hyperplane* adalah garis atau bidang yang memisahkan data dari dua kelas dalam ruang fitur. Dalam konteks SVM, *hyperplane* optimal adalah yang memaksimalkan margin, yaitu jarak antara *hyperplane* dengan data terdekat dari masing-masing kelas. Margin yang lebih besar menunjukkan generalisasi yang lebih baik dari model SVM. Persamaan *hyperplane* untuk memisahkan dua kelas dengan margin terbesar ditulis sebagai berikut:

$$w.x + b = 0 \quad (1)$$

Vektor bobot sebagai penunjuk arah hyperplane direpresentasikan ke dalam w, sementara x adalah vektor input, dan b adalah bias atau konstanta *offset*. *Hyperplane* optimal memiliki margin maksimum untuk memastikan generalisasi model terhadap data baru.

Tujuan SVM adalah memaksimalkan margin, yaitu jarak antara *hyperplane* dengan data terdekat dari kedua kelas. Fungsi optimasinya dinyatakan sebagai:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (2)$$

Rumus ini memastikan bahwa semua data berada di sisi *hyperplane* yang benar dan margin terbesar tercapai.

Untuk data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk memetakan data ke dimensi yang lebih tinggi. Fungsi kernel ini memungkinkan SVM menangani data yang kompleks dan memiliki distribusi non-linear. Salah satu jenis kernel yang populer adalah kernel *Gaussian* atau RBF *(Radial Basis Function)*, yang mengukur kesamaan antara dua titik dalam ruang fitur (Elen dkk, 2022) (Angel Deborah dkk, 2021).

Keunggulan utama SVM adalah kemampuannya menangani data set berdimensi tinggi. Hal ini disebabkan SVM hanya memanfaatkan data yang berada di tepi margin *(support vectors)* sehingga, algoritma ini efisien dalam penggunaan sumber daya komputasi. Selain itu, SVM memiliki regularisasi bawaan yang meminimalkan risiko *overfitting*, bahkan pada data set dengan jumlah data yang lebih kecil dibandingkan jumlah fitur. Namun, SVM sangat bergantung pada representasi data yang baik. Jika fitur yang digunakan tidak relevan atau tidak cukup informatif, performa SVM akan menurun. Pemilihan fungsi kernel dan parameter yang tepat juga menjadi tantangan yang signifikan dalam penerapannya (Kurani dkk, 2023).

Dua metode utama yang digunakan untuk merepresentasikan data teks dalam SVM adalah *CountVectorizer* dan TF-IDF. *CountVectorizer* adalah pendekatan sederhana yang merepresentasikan teks sebagai vektor berdasarkan jumlah kemunculan kata dalam dokumen. Setiap dokumen direpresentasikan sebagai *array* nilai, yaitu setiap elemen menunjukkan frekuensi kata tertentu dalam dokumen tersebut (Krabokoukis & Polyzos, 2023).

Sebagai alternatif, TF-IDF *(Term Frequency-Inverse Document Frequency)* memberikan bobot lebih tinggi pada kata-kata unik yang memiliki nilai

informasi lebih tinggi dalam dokumen tertentu. TF-IDF menggabungkan dua metrik yaitu *Term Frequency* (TF), yang menghitung frekuensi kemunculan kata dalam dokumen, dan *Inverse Document Frequency* (IDF), yang mengukur seberapa unik kata tersebut di seluruh dokumen. Pendekatan ini membantu mengurangi pengaruh kata-kata umum yang sering muncul di semua dokumen, seperti kata fungsional "dan" atau "dengan" (Bakiyev, 2022).

Sebagai contoh, jika sebuah dokumen memiliki kata-kata "informasi" yang sering muncul di semua dokumen, *CountVectorizer* akan memberikan bobot yang sama seperti kata "analisis" yang lebih jarang muncul tetapi penting untuk kategori tertentu. TF-IDF akan memberikan bobot lebih tinggi pada "analisis," sehingga model lebih peka terhadap konteks dokumen.

Dalam konteks klasifikasi kalimat baku dan tidak baku, TF-IDF menjadi alat yang sangat penting untuk meningkatkan sensitivitas model terhadap pola-pola unik yang ada dalam kedua jenis kalimat tersebut. Dengan memberikan bobot yang lebih akurat pada kata-kata kunci, TF-IDF memungkinkan SVM untuk lebih efektif membedakan antara kalimat baku dan tidak baku. Kombinasi antara SVM yang *robust* dan representasi data yang relevan seperti TF-IDF memberikan solusi yang efisien untuk berbagai tugas klasifikasi teks.

2.3. Augmentasi Data

Augmentasi Augmentasi data merupakan teknik yang bertujuan untuk meningkatkan ukuran dan keragaman data set dengan menciptakan data baru berdasarkan data asli. Dalam pengolahan teks, teknik ini memainkan peran penting dalam membantu model pembelajaran mesin memahami pola yang lebih kompleks atau menangani data yang tidak seimbang. Berbagai pendekatan dapat digunakan untuk melakukan augmentasi data, seperti mengganti kata-kata tertentu dengan sinonim untuk menciptakan variasi tanpa mengubah makna asli, melakukan parafrase untuk mengubah struktur kalimat tanpa mengubah konteks, atau menyisipkan dan menghapus kata-kata guna memperkaya struktur teks. Teknik ini secara efektif memperluas cakupan pola pelatihan model, sehingga dapat meningkatkan generalisasi pada data baru (Shorten dkk, 2021).

Dalam penelitian ini, augmentasi data difokuskan pada kalimat baku dengan menambahkan contoh-contoh yang mencerminkan pola ambigu, seperti kutipan atau ungkapan puitis. Kalimat baku sering kali memiliki struktur yang kompleks dan sulit dibedakan dari kalimat tidak baku ketika menggunakan pendekatan berbasis fitur sederhana. Dengan menambahkan variasi teks yang menyerupai pola-pola ini, augmentasi data membantu model belajar untuk mengenali ciri khas kalimat baku sekaligus mengurangi kesalahan klasifikasi. Pendekatan ini bertujuan untuk memperbaiki kelemahan model sebelumnya yang sering salah mengklasifikasikan kalimat baku sebagai tidak baku (Sujana & Kao, 2023).

3. METODE PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk mengevaluasi performa model SVM dalam klasifikasi kalimat baku dan tidak baku berbahasa Indonesia. Data yang digunakan terdiri dari 2.000 kalimat berbahasa Indonesia yang telah diberi label baku dan tidak baku. Data diperoleh dari berbagai sumber, termasuk dokumen resmi, artikel berita, media sosial, dan percakapan sehari-hari. Data set ini telah melalui tahap pra-pemrosesan seperti penghapusan tanda baca yang tidak relevan, normalisasi teks, serta penghilangan angka atau simbol yang tidak bermakna. Contoh data set yang dipergunakan dalam penelitian ditampilkan dalam Tabel 1.

Tabel 1. Contoh Kalimat Baku dan Tidak Baku

Label	Kalimat
Baku	(1) Dokumen ini harus diserahkan sebelum tanggal 20 Desember 2024. (2) Makalah ini bertujuan untuk mengetahui sentimen masyarakat terhadap tindakan vaksinasi. (3) Kombinasi kedua metode tersebut juga dapat meminimalkan *noise region* yang mengakibatkan kesalahan dalam proses deteksi dan pelacakan. (4) Pneumonia menjadi masalah kesehatan yang utama di Indonesia. (5) Candi Prambanan termasuk situs warisan dunia UNESCO yang merupakan candi hindu terbesar di Indonesia
Tidak baku	(1) Bro, gua udah kirim file-nya tadi siang. (2) Bercanda tapi lo kek mau nelen gue bulet ih. (3) Konten serius dibercandain konten bercanda diseriusin emang aneh manusia. (4) Bercanda ayangku cintaku manisku. (5) Belum bisa bedain marah beneran atau cuma sekedar bercanda yg pake bahasa kasar.

Data set ini dibagi menggunakan pembagian khusus yang mempertimbangkan rasio representatif antara data latih dan data uji, yaitu 70% data untuk pelatihan, 15% untuk validasi, dan 15% untuk pengujian. Rasio ini dipilih untuk memastikan keseimbangan dalam evaluasi performa model (Nguyen dkk, 2021).

3.2. Tahapan Penelitian

Tahapan penelitian dimulai dengan penyusunan data set yang melibatkan pengumpulan dan pra-pemrosesan data untuk memastikan kualitas input model. Eksperimen utama dilakukan untuk melatih model SVM menggunakan kernel linier, dengan evaluasi kinerja menggunakan metrik akurasi, *precision, recall*, dan *F1-score*. Evaluasi awal, data

uji digunakan untuk mengukur performa model secara langsung setelah pelatihan.

Untuk meningkatkan performa model, dilakukan augmentasi data dengan menambahkan variasi teks melalui:

1. Penambahan sinonim untuk kata-kata umum dalam teks.
2. Penyusunan ulang frasa tanpa mengubah makna teks.
3. Penambahan kalimat baru yang mencerminkan pola ambigu seperti kalimat puitis atau kutipan.

Sebagai bagian dari strategi perbaikan, metode *CountVectorizer* digantikan dengan TF-IDF *Vectorizer*. Penggunaan TF-IDF membantu memberikan bobot lebih besar pada kata-kata unik yang relevan dalam membedakan kalimat baku dan tidak baku. Penelitian juga memanfaatkan Python dengan pustaka Scikit-Learn, Pandas, NumPy, dan Matplotlib untuk analisis serta visualisasi data, sementara augmentasi teks dilakukan menggunakan pustaka NLP seperti NLTK dan SpaCy. Eksperimen sepenuhnya dilakukan pada *platform Google Colab* untuk memanfaatkan akses komputasi berbasis *cloud* yang efisien.

3.3. Analisis Data

Analisis kesalahan dilakukan untuk mengidentifikasi pola umum pada prediksi yang salah. Semua kesalahan prediksi ditemukan pada kalimat baku yang salah diklasifikasikan sebagai tidak baku. Rata-rata panjang teks salah prediksi adalah 80.67 karakter, dibandingkan dengan 109.35 karakter untuk teks yang diklasifikasikan benar. Kesalahan ini sebagian besar terjadi pada teks bergaya puitis atau kutipan yang memiliki sifat ambigu.

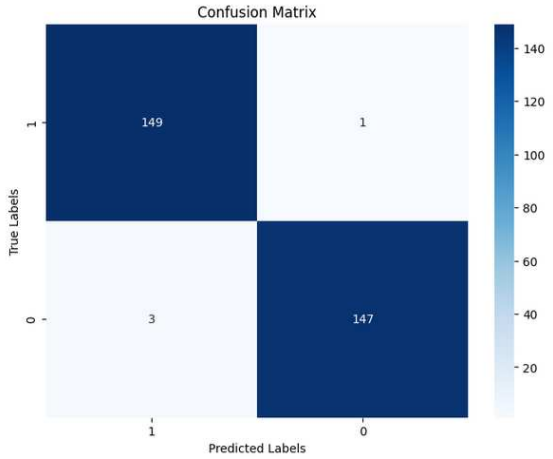
Validasi model dilakukan berdasarkan hasil evaluasi langsung pada data uji, tanpa menggunakan teknik *k-fold cross-validation*. Keputusan ini diambil karena penelitian bersifat *proof of concept* dengan fokus utama pada pola kesalahan spesifik, serta mempertimbangkan keterbatasan *dataset* dan efisiensi komputasi. Hal ini disesuaikan dengan tujuan penelitian untuk fokus pada pola kesalahan spesifik daripada generalisasi melalui pembagian data yang berulang. Evaluasi model menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk memberikan gambaran menyeluruh tentang kemampuan model.

Metodologi ini dirancang untuk memberikan hasil penelitian yang relevan dalam konteks klasifikasi kalimat baku dan tidak baku berbahasa Indonesia. Jika diperlukan pengembangan lebih lanjut, penambahan data atau penggunaan model yang lebih kompleks serta penerapan *k-fold cross-validation* dapat dipertimbangkan.

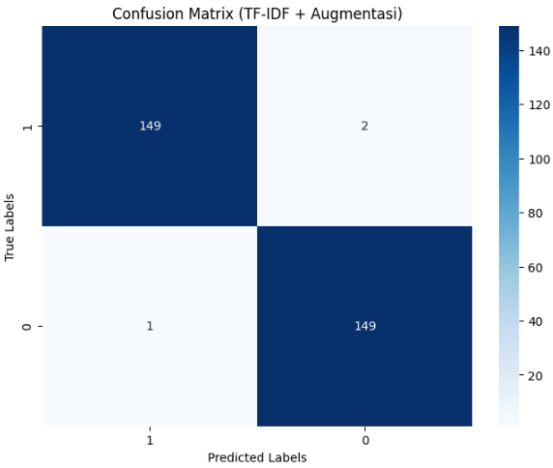
4. HASIL DAN DISKUSI

Penelitian ini bertujuan untuk mengevaluasi performa model *Support Vector Machine* (SVM) dalam klasifikasi kalimat baku dan tidak baku berbahasa Indonesia. Model dilatih menggunakan *Countvectorizer* dan TF-IDF *Vectorizer* untuk menghasilkan representasi kalimat dalam bentuk numerik.

Hasil pengujian menunjukkan bahwa model mencapai akurasi sebesar 99.3% setelah augmentasi data diterapkan. Sebelum augmentasi, model memiliki akurasi sebesar 98.6%, dengan beberapa kesalahan prediksi yang lebih signifikan pada kalimat baku. Untuk membandingkan perbaikan hasil klasifikasi, berikut adalah *confusion matrix* (CM) sebelum dan sesudah augmentasi yang ditampilkan pada Gambar 1 dan Gambar 2.



Gambar 1. *Confusion Matrix* pada *Dataset* Asli



Gambar 2. *Confusion Matrix* pada *Dataset* setelah Augmentasi

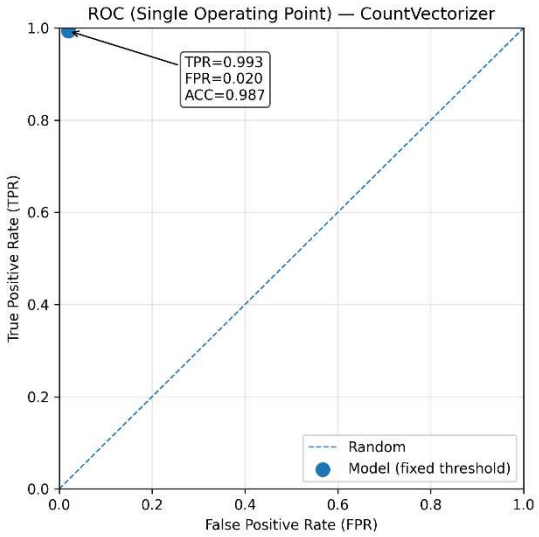
Secara rinci data yang dimuat dalam *confusion matrix* ditampilkan dalam Tabel 2.

Tabel 2. Perbandingan sebelum dan sesudah augmentasi

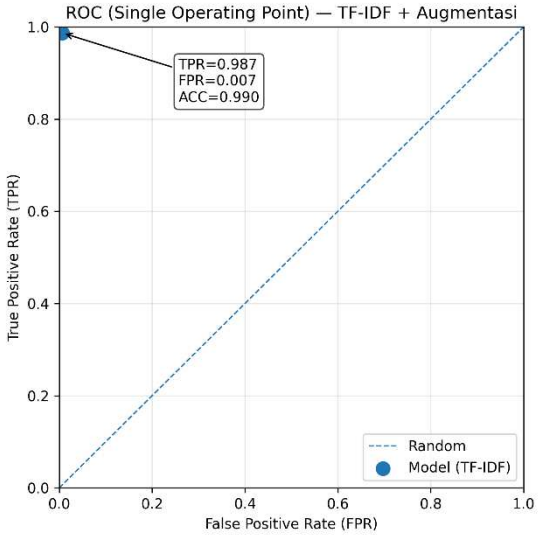
Kondisi	TP	FN	TN	FP	Kesalahan
Sebelum	149	1	147	3	4
Sesudah	149	2	149	1	3

Selain *confusion matrix* dan *classification report*, analisis kinerja model juga diperkuat dengan

visualisasi *Receiver Operating Characteristic* (ROC). Pada metode *CountVectorizer* yang ditampilkan dalam Gambar 3, ROC menunjukkan titik kinerja model dengan tingkat akurasi 98.6%, sedangkan pada metode TF-IDF dengan augmentasi yang ditampilkan pada Gambar 4, posisi titik ROC semakin dekat dengan area optimal dengan akurasi 99.3%. Visualisasi ini menegaskan bahwa perbaikan representasi data dan augmentasi berkontribusi nyata dalam menurunkan jumlah kesalahan klasifikasi.



Gambar 3. Visualisasi ROC *CountVectorizer*



Gambar 4. Visualisasi ROC TF-IDF Augmentasi

Perbedaan data pada kedua *confusion matrix* yang ditampilkan dalam Gambar 1 dan Gambar 2, menunjukkan bahwa augmentasi data memberikan kontribusi signifikan dalam meningkatkan sensitivitas model terhadap kalimat baku. Perbaikan terbesar terjadi pada pengurangan jumlah kesalahan *False Positives*, dari 3 menjadi 1, yang mencerminkan peningkatan pemahaman model terhadap pola unik dalam kalimat tidak baku. TF-IDF *Vectorizer* memainkan peran penting dalam memberikan bobot lebih pada kata-kata yang informatif, sehingga membantu model membedakan pola antara kalimat baku dan tidak baku dengan lebih baik. Augmentasi data juga memperluas cakupan pelatihan, sehingga memungkinkan model menangani variasi kalimat baku yang lebih kompleks.

Namun, peningkatan *False Negatives* dari 1 menjadi 2 mengindikasikan bahwa pola ambigu dalam kalimat baku, seperti elemen puitis atau gaya bahasa informal, masih menjadi tantangan. Hal ini menunjukkan bahwa augmentasi data manual belum sepenuhnya mencakup pola-pola kompleks tersebut. Dalam konteks aplikasi nyata, pengurangan *False Positives* sangat penting untuk memastikan integritas data pada analisis dokumen resmi, sementara *False Negatives* memerlukan perhatian lebih terutama dalam menjaga kualitas respons pada *chatbot*.

Dibandingkan dengan pendekatan sebelumnya, hasil ini menunjukkan efektivitas kombinasi TF-IDF dan augmentasi manual, meskipun skalabilitas metode ini tetap menjadi keterbatasan yang perlu diatasi pada penelitian mendatang.

Kesalahan prediksi yang tersisa sebagian besar terjadi pada kalimat baku yang pendek atau ambigu dengan struktur sederhana serta penggunaan kata-kata umum. Untuk menggambarkan pola tersebut, berikut ditampilkan beberapa contoh ilustratif yang cenderung sulit diklasifikasikan dengan benar oleh model:

1. Waktu adalah segalanya.
2. Hidup itu indah.
3. Damai itu indah namun sulit dijaga.
4. Perjuangan itu tak akan sia-sia.
5. Kesuksesan datang dari kerja keras.
6. Pendidikan adalah kunci masa depan.
7. Kebahagiaan sejati tidak bisa dibeli.
8. Hidup adalah perjalanan penuh misteri.
9. Kejujuran adalah dasar kepercayaan.
10. Waktu tak akan kembali, gunakan sebaik-baiknya.

Kalimat-kalimat seperti "Waktu adalah segalanya" dan "Hidup itu indah" memiliki nuansa puitis dan filosofis yang membuatnya sulit diklasifikasikan oleh model pembelajaran mesin. Nuansa ini disebabkan oleh sifat abstrak dari kata-kata yang digunakan, seperti "indah" dan "waktu," yang tidak secara eksplisit menunjukkan konteks baku atau tidak baku. Model pembelajaran mesin, seperti *Support Vector Machine* (SVM), mengandalkan pola dan fitur spesifik untuk membuat prediksi.

Namun, dalam kasus kalimat ini, pola yang dihasilkan sering kali ambigu karena karakteristiknya dapat ditemukan baik dalam kalimat baku maupun tidak baku. Hal ini semakin diperumit oleh fakta bahwa struktur pendek dari kalimat-kalimat ini mengurangi jumlah informasi tekstual yang dapat dimanfaatkan oleh model untuk melakukan klasifikasi yang akurat.

Ambiguitas dalam konteks penggunaan kata juga menjadi faktor yang signifikan. Kalimat seperti "Damai itu indah namun sulit dijaga" atau "Perjuangan itu tak akan sia-sia" dapat digunakan dalam berbagai situasi, baik baku maupun tidak baku, tergantung pada pengirim dan penerima pesan. Dalam konteks percakapan sehari-hari, kalimat seperti ini mungkin dianggap tidak baku karena sering digunakan untuk mengekspresikan pendapat pribadi secara santai.

Namun, dalam dokumen resmi atau presentasi formal, kalimat-kalimat tersebut dapat dianggap baku karena sesuai dengan norma penggunaan bahasa baku. Kesamaan konteks semantik ini menyulitkan model untuk menentukan klasifikasi yang tepat tanpa adanya fitur tambahan atau konteks yang lebih dalam.

Lebih jauh lagi, kata-kata yang digunakan dalam kalimat ini bersifat netral dan sering ditemukan dalam kalimat baku maupun tidak baku. Sebagai contoh, kata "indah," "waktu," atau "hidup" memiliki makna yang umum, sehingga sulit untuk menentukan konteks hanya dari kehadiran kata-kata tersebut.

Pola semantik yang terbatas ini menunjukkan bahwa pendekatan berbasis fitur sederhana, seperti *CountVectorizer* atau bahkan TF-IDF *Vectorizer*, memiliki keterbatasan dalam menangani kasus seperti ini. Oleh karena itu, augmentasi data yang lebih terarah, khususnya dengan menambahkan variasi kalimat baku pendek dengan gaya ambigu, sangat diperlukan untuk melatih model agar mampu menangkap perbedaan halus antara kalimat baku dan tidak baku. Pendekatan ini juga harus dikombinasikan dengan teknik pemodelan berbasis konteks untuk meningkatkan sensitivitas terhadap pola ambigu.

Untuk menjawab tantangan ini, augmentasi data dilakukan secara manual dengan campur tangan manusia untuk menjaga kualitas dan relevansi teks. Sebanyak 50 kalimat baku yang ambigu dipilih dan diaugmentasi menggunakan pendekatan sebagai berikut:

1. Penggantian Sinonim yaitu kata-kata umum dalam kalimat diganti dengan sinonimnya untuk mempertahankan makna tetapi mengubah strukturnya. Contoh dalam proses ini adalah:
    a. Sebelum: "Jangan berhenti mencoba sebelum berhasil."
    b. Sesudah: "Jangan pernah menyerah sebelum mencapai kesuksesan."
2. Penyusunan ulang frasa yaitu struktur kalimat disusun ulang agar variasi data lebih beragam. Contoh dalam proses ini adalah:
    a. Sebelum: "Hidup adalah perjalanan yang penuh misteri."
    b. Sesudah: "Perjalanan hidup selalu menyimpan misteri yang mendalam."

Hasil augmentasi ini membantu model memahami variasi kalimat baku yang lebih kompleks, khususnya kalimat ambigu atau puitis yang cenderung diklasifikasikan salah. Peningkatan akurasi dari 98.6% sebelum augmentasi menjadi 99.3% setelah augmentasi, meskipun kecil, menunjukkan perbaikan signifikan dalam mengatasi *False Positives*.

Meskipun model berbasis *Deep Learning* sering menjadi pilihan utama dalam pemrosesan bahasa alami, hasil penelitian ini menunjukkan bahwa SVM masih dapat dimanfaatkan secara efektif, khususnya dalam kondisi berikut:

1. Data set yang Terbatas: SVM bekerja dengan baik pada data set berukuran kecil hingga menengah, seperti yang digunakan dalam penelitian ini, karena kompleksitas komputasinya lebih rendah dibandingkan model *deep learning*.
2. Representasi Fitur Sederhana: Dengan menggunakan TF-IDF, SVM mampu menangkap pola teks dengan baik tanpa memerlukan representasi konteks yang lebih dalam seperti pada model transformer.
3. Ambiguitas Teks Terkendali: Dalam kasus kalimat ambigu, augmentasi data sederhana mampu membantu meningkatkan performa SVM dengan biaya komputasi yang lebih ringan.

Peningkatan ini terutama disebabkan oleh augmentasi manual yang memberikan variasi pada pola kalimat baku ambigu. Namun, augmentasi manual ini juga menjadi keterbatasan dalam penelitian karena tidak praktis jika diterapkan pada skala data yang lebih besar. Selain itu, jumlah data yang digunakan dalam penelitian ini relatif terbatas (2.000 kalimat), sehingga hasil yang diperoleh perlu dipandang sebagai indikasi awal yang masih menyisakan ruang untuk generalisasi yang lebih luas. Dalam skenario tersebut, pendekatan augmentasi otomatis berbasis NLP, seperti penggunaan model transformer atau algoritma generatif, menjadi solusi potensial untuk masa depan.

Penelitian ini juga menegaskan bahwa kombinasi pendekatan berbasis fitur sederhana dengan augmentasi data dapat memberikan hasil yang optimal dalam klasifikasi kalimat baku dan tidak baku. Dengan demikian, penelitian ini memberikan kontribusi praktis untuk aplikasi seperti analisis dokumen resmi, pengelompokan teks, dan sistem klasifikasi yang peka terhadap variasi gaya bahasa di Indonesia.

Arah pengembangan penelitian selanjutnya mencakup eksplorasi model berbasis *deep learning* untuk menangkap konteks yang lebih dalam, serta penggunaan augmentasi otomatis untuk meningkatkan efisiensi dalam skala data yang lebih luas.

5. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa *Support Vector Machine* (SVM) dapat mencapai

performa yang sangat baik dalam klasifikasi kalimat baku dan tidak baku berbahasa Indonesia dengan akurasi 99.3%, terutama ketika dikombinasikan dengan TF-IDF *Vectorizer* dan augmentasi data. Pendekatan ini membuktikan bahwa SVM tetap efektif dan memberikan hasil optimal dengan efisiensi komputasi dibandingkan model yang lebih kompleks. Penggunaan TF-IDF berhasil mengurangi *False Positives* secara signifikan, meskipun *False Negatives* sedikit meningkat. Hasil ini menunjukkan bahwa SVM sesuai untuk menangani klasifikasi kalimat dalam skenario terbatas. Untuk mengatasi keterbatasan augmentasi manual, penelitian selanjutnya dapat memanfaatkan metode augmentasi otomatis berbasis *Natural Language Processing* (NLP), seperti transformer atau algoritma generatif, yang dapat mendukung pendekatan berbasis SVM.

DAFTAR PUSTAKA

ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., LI, J. dan NARAG, M., 2021. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. Artificial Intelligence in Medicine, 120, p.102167.

ANGEL DEBORAH, S., MIRNALINEE, T.T. dan RAJENDRAM, S.M., 2021. Emotion Analysis on Text Using Multiple Kernel Gaussian. Neural Processing Letters, 53(2), pp.1187–1203.

BAKIYEV, B., 2022. Method for Determining the Similarity of Text Documents for the Kazakh language, Taking Into Account Synonyms: Extension to TF-IDF. SIST 2022 - 2022 International Conference on Smart Information Systems and Technologies, Proceedings.

BAYER, M., KAUFHOLD, M.A., BUCHHOLD, B., KELLER, M., DALLMEYER, J. dan REUTER, C., 2023. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International Journal of Machine Learning and Cybernetics, 14(1), pp.135–150.

BAYER, M., KAUFHOLD, M.A. dan REUTER, C., 2022. A Survey on Data Augmentation for Text Classification. ACM Computing Surveys, 55(7).

BENGESI, S., OLADUNNI, T., OLUSEGUN, R. dan AUDU, H., 2023. A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets. IEEE Access, 11, pp.11811–11826.

BILGEN, Y. dan KAYA, M., 2024. EGMA: Ensemble Learning-Based Hybrid Model Approach for Spam Detection. Applied Sciences 2024, 14(21), p.9669.

DANYAL, M.M., KHAN, S.S., KHAN, M., ULLAH, S., GHAFFAR, M.B. dan KHAN, W., 2024. Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer. Social Network Analysis and Mining, 14(1), pp.1–15.

ELEN, A., BAŞ, S. dan KÖZKURT, C., 2022. An Adaptive Gaussian Kernel for Support Vector Machine. Arabian Journal for Science and Engineering, 47(8), pp.10579–10588.

KHAIRUL, M.F.R. dan PERDANA, R.S., 2024. Arsitektur Sistem Percakapan Otomatis Berbahasa Indonesia dengan Normalisasi Bahasa Informal Menjadi Baku. Jurnal Teknologi Informasi dan Ilmu Komputer, [online] 11(5), pp.1009–1016. https://doi.org/10.25126/JTIIK.2024117984.

GABRIELA, N.H., SIAUTAMA, R., AMADEA, C.I.A. dan SUHARTONO, D., 2021. Extractive Hotel Review Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends. Procedia Computer Science, 179, pp.558–565.

GOYAL, R., 2021. Evaluation of rule-based, CountVectorizer, and Word2Vec machine learning models for tweet analysis to improve disaster relief. 2021 11th IEEE Global Humanitarian Technology Conference, GHTC 2021, pp.16–19.

HAITAO, W., JIE, H., XIAOHONG, Z. dan SHUFEN, L., 2020. A Short Text Classification Method Based on N-Gram and CNN. Chinese Journal of Electronics, 29(2), pp.248–254.

HANA, K.M., ADIWIJAYA, AL FARABY, S. dan BRAMANTORO, A., 2020. Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines. 2020 International Conference on Data Science and Its Applications, ICoDSA 2020.

KRABOKOUKIS, T. dan POLYZOS, S., 2023. A Bibliometric Analysis of Integrating Tourism Development into Urban Planning. Sustainability 2023, 15(20), p.14886.

KURANI, A., DOSHI, P., VAKHARIA, A. dan SHAH, M., 2023. A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting. Annals of Data Science, 10(1), pp.183–208.

LIU, H., CHEN, X. dan LIU, X., 2022. A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis. IEEE Access, 10, pp.32280–32289.

NGUYEN, Q.H., LY, H.B., HO, L.S., AL-ANSARI, N., VAN LE, H., TRAN, V.Q., PRAKASH, I. dan PHAM, B.T., 2021. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering, 2021(1), p.4832864.

NURLIZA, N.S., HIDAYAH, N., AZZAHRA, S.V. dan GINANJAR, B., 2025. Afiks Ng- pada Bahasa Gaul di Media Sosial Beserta Padanan Formalnya: Kajian Morfologi. Linguistik Indonesia, [online] 43(1), pp.81–98. https://doi.org/10.26499/li.v43i1.673.

QIU, S., XU, B., ZHANG, J., WANG, Y., SHEN, X., DE MELO, G., LONG, C. dan LI, X., 2020. EasyAug: An Automatic Textual Data Augmentation Platform for Classification Tasks. The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020, pp.249–252.

RIANTO, BENNY MUTIARA, A., PRASETYO WIBOWO, E. dan INSAP SANTOSA, P., 2020. The crowdsourcing method to normalize 'bahasa alay', a case of indonesian corpus. 2020 5th International Conference on Informatics and Computing, ICIC 2020.

RIANTO, MUTIARA, A.B., WIBOWO, E.P. dan SANTOSA, P.I., 2021. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. Journal of Big Data, 8(1), pp.1–16.

SHORTEN, C., KHOSHGOFTAAR, T.M. dan FURHT, B., 2021. Text Data Augmentation for Deep Learning. Journal of Big Data, 8(1), pp.1–34.

SUJANA, Y. dan KAO, H.Y., 2023. LiDA: Language-Independent Data Augmentation for Text Classification. IEEE Access, 11, pp.10894–10901.

TRILOKA, J., HARTONO, H. dan SUTEDI, S., 2022. Detection of SQL Injection Attack Using Machine Learning Based On Natural Language Processing. International Journal of Artificial Intelligence Research, [online] 6(2). https://doi.org/10.29099/ijair.v6i2.355.

Halaman ini sengaja dikosongkan