The Impact of Text Data Preprocessing for Review Analysis E-Wallet Application on Google Play Store [Pengaruh Preprocessing Data Teks untuk Analisis Ulasan Aplikasi E-Wallet di Google Play Store]

Arrizqi Fauzy Aufar[1)], Mochamad Alfan Rosid*[,2)], Ade Eviyanti*[3)], Ika Ratna Indra Astutik*[4)]

[1)] Program Studi Infornatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2)] Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[3)] Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[4)] Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: alfanrosid@umsida.ac.id, adeeviyanti@umsida.ac.id, ikaratna@umsida.ac.id

Abstract. *This Research aims to optimize preprocessing techniques in sentiment analysis of reviews for the E-Wallet Dana application on the Google Play Store. Text preprocessing is a crucial step in Natural Language Processing (NLP) that affects the accuracy and efficiency of sentiment analysis. This study employs various preprocessing methods, including stopwords removal, stemming, and lemmatization, to clean and prepare the review data before analysis. The results show that lemmatization techniques significantly improve accuracy compared to basic preprocessing techniques such as stopwords removal and stemming. With proper preprocessing optimization, sentiment analysis can provide more accurate and informative results, which can be used to enhance the application's quality and user experience. This study uses SVM classification testing models with 4 kernels, where the highest results were achieved with cleaning, case folding, tokenization, and lemmatization techniques at 100% for Linear; 100% for RBF, 99% for Polynomial, and 99.50% for Sigmoid with an average accuracy of 99.63%.*

Keywords - *DANA; Google Play Store; Preprocessing; Sentiment Analysis*

Abstrak. *Penelitian ini bertujuan untuk mengoptimalkan teknik preprocessing dalam analisis sentimen ulasan aplikasi E-Wallet Dana di Google Play Store. Preprocessing teks merupakan langkah penting dalam pemrosesan bahasa alami (Natural Language Processing/NLP) yang mempengaruhi akurasi dan efisiensi analisis sentimen. Studi ini menggunakan berbagai metode preprocessing, termasuk penghapusan stopwords, stemming, dan lemmatization untuk membersihkan dan mempersiapkan data ulasan sebelum dianalisis. Hasil penelitian menunjukkan bahwa teknik lemmatization memberikan peningkatan akurasi yang signifikan dibandingkan dengan teknik preprocessing dasar seperti penghapusan stopwords dan stemming. Dengan optimasi preprocessing yang tepat, analisis sentimen dapat memberikan hasil yang lebih akurat dan informatif, yang dapat digunakan untuk meningkatkan kualitas aplikasi dan pengalaman pengguna. Dalam penelitian ini menggunakan model pengujian klasifikasi SVM 4 kernel dimana hasil tertinggi pada teknik cleaning, case folding, tokenization, dan Lemmatization pada 100% Linear; 100% RBF, 99% Polynomial, dan 99,50% Sigmoid dengan rata – rata 99,63%.*

Kata Kunci – *Analisis Sentimen; DANA; Google Play Store; Preprocessing*

I. PENDAHULUAN

Dalam era digitalisasi yang berkembang pesat ini, layanan E-Wallet telah menjadi bagian yang tak terpisahkan dari kehidupan sehari-hari pengguna di berbagai negara. Kemajuan teknologi keuangan, kenyamanan, dan kemudahan akses yang ditawarkan oleh *E-Wallets* telah memicu adopsi yang signifikan dari solusi pembayaran ini. E-wallet, atau dompet elektronik adalah aplikasi atau perangkat lunak yang memungkinkan pengguna untuk melakukan transaksi keuangan secara digital. Ini termasuk menyimpan uang, melakukan pembayaran, dan menerima uang melalui ponsel atau perangkat lain yang terhubung ke internet. E-wallet sering digunakan untuk pembelian online, transfer uang, dan pembayaran di toko-toko fisik menggunakan kode QR atau teknologi nirkontak lainnya. Contoh e-wallet yang populer di Indonesia adalah OVO, GoPay, DANA, dan LinkAja[1].

DANA adalah salah satu *e-wallet* yang populer di Indonesia karena berbagai alasan. Antarmuka yang *user-friendly* membuat aplikasi ini mudah digunakan oleh berbagai kalangan, dari remaja hingga orang dewasa. DANA menawarkan fitur keamanan yang kuat, termasuk otentikasi dua faktor dan enkripsi data, untuk melindungi transaksi pengguna. Selain itu, DANA memungkinkan berbagai jenis transaksi, seperti pembayaran tagihan, pembelian pulsa, transfer uang, dan belanja online serta offline[2]. Pengguna juga tertarik dengan promosi, cashback, dan diskon yang sering ditawarkan oleh DANA. Kolaborasi dan integrasi dengan berbagai merchant, platform e-commerce, dan bank meningkatkan kemudahan dan fleksibilitas dalam bertransaksi. Didukung oleh EMTEK Group dan *Ant Financial* (bagian dari Alibaba Group), DANA memiliki dukungan teknologi dan sumber daya yang kuat, memastikan

kehandalan dan inovasi dalam layanan yang diberikan. Alasan-alasan ini menjadikan DANA pilihan populer di kalangan pengguna *e-wallet* di Indonesia[3].

Perkembangan era digital telah menghasilkan dinamika baru dalam kaitannya dengan Keterkaitan antara penyedia layanan E-Wallet dan pengguna. Pengguna E-Wallet Dana, seperti pengguna layanan serupa, semakin sering menyuarakan pengalaman mereka melalui ulasan dan komentar di platform distribusi seperti Google Play Store. Ulasan ini bukan hanya menjadi sumber berharga bagi calon pengguna yang ingin mengevaluasi aplikasi Dana, tetapi juga merupakan informasi berharga bagi perusahaan itu sendiri. Menganalisis sentimen dalam ulasan pengguna dapat memberikan wawasan yang dalam tentang sejauh mana pengguna puas atau tidak puas dengan layanan Dana, serta menyoroti masalah yang perlu diperbaiki[2].

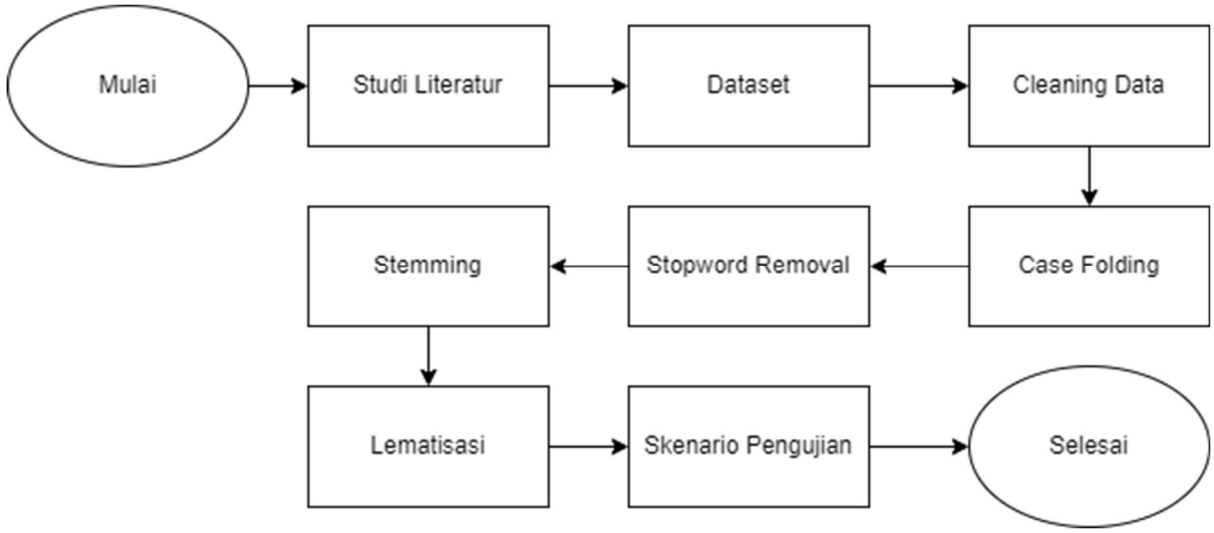
Banyaknya ulasan aplikasi E-Wallet Dana di Google Play Store seringkali disertai dengan masalah seperti kesalahan ejaan, penggunaan slang, atau informasi yang tidak relevan, yang dapat mengganggu analisis sentimen yang akurat. Proses preprocessing data teks untuk analisis sentimen memerlukan pemilihan teknik yang tepat untuk membersihkan dan mempersiapkan data dengan benar. Namun, banyaknya variasi dalam gaya penulisan dan bahasa di ulasan pengguna dapat menyulitkan pemilihan teknik yang sesuai. Seperti dalam jurnal yang ditulis oleh Mochamad Alfan Rosid bahwa dalam pemilihan metode *Preprocessing Teks* sangat mempengaruhi dalam menentukan keakurasian. Penelitian itu membandingan pada metode *Porter Stemmer* dengan *Sastrawi*[4].

Penelitian Rama Ulgasesa meneliti tentang pengaruh *stemming* pada teks preposeing untuk mengetahui performa keakuratan pada sentimen analisis tentang kebijakan New Normal pada aplikasi Twitter. Penelitian ini bertujuan untuk membandingkan kinerja klasifikasi antara menggunakan proses stemming dan tidak menggunakan stemming pada dataset melalui proses preprocessing, serta untuk mengidentifikasi algoritma klasifikasi yang memberikan hasil terbaik ketika stemming diterapkan dalam tahapan preprocessing[5].

Penelitian ini bertujuan agar dapat meningkatkan akurasi dalam menangkap sentimen pengguna terhadap aplikasi Dana dengan mengatasi masalah-masalah preprocessing yang ada, sehingga memberikan wawasan yang lebih akurat kepada penyedia layanan. hasil penelitian ini akan memiliki dampak positif pada pengembangan produk dan layanan Dana, membantu mereka dalam meningkatkan kepuasan pelanggan dan menjaga kompetitivitas di pasar digital.

II. METODE

Pada bab Metodologi ini akan menggunakan beberapa tahapan atau alur penelitian agar mendapatkan hasil yang diinginkan, bisa dilihat pada gambar 1.



Gambar 1 Alur Penelitian

A. Studi literatur

Dalam fase Studi Literatur, peneliti mengumpulkan referensi dari berbagai sumber media, seperti buku, jurnal akademik, dan e-book untuk mendukung penelitian mereka. Proses ini bertujuan untuk mendapatkan pemahaman yang menyeluruh tentang teori dasar yang relevan dan hasil-hasil penelitian yang telah dilakukan oleh peneliti sebelumnya. Dengan menganalisis literatur yang ada, peneliti dapat membangun landasan teoritis yang kuat untuk penelitian mereka. Studi literatur ini memungkinkan peneliti untuk mengeksplorasi berbagai perspektif dan pendekatan yang telah digunakan dalam topik yang sama, serta mengidentifikasi tren, temuan, dan metodologi yang signifikan. Selain itu, fase ini membantu dalam mengidentifikasi kesenjangan dalam penelitian yang ada, yaitu area-area yang belum

sepenuhnya dieksplorasi atau yang masih memerlukan penyelidikan lebih lanjut. Dengan mengisi kesenjangan ini, penelitian saat ini dapat memberikan kontribusi yang berarti dan memperluas pengetahuan yang ada. Proses ini juga memastikan bahwa penelitian yang dilakukan tidak hanya relevan dan berlandaskan pada teori yang solid, tetapi juga inovatif dan memberikan nilai tambah pada bidang studi yang sedang diteliti.

B. Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 1000 ulasan yang diperoleh melalui proses pengikisan menggunakan API *Google Play Scraper*. Pengumpulan data ini dilakukan pada rentanbulan januari sampai maret tahun 2024 untuk memastikan bahwa data yang diperoleh adalah yang paling relevan dan terkini. Ulasan-ulasan yang dikumpulkan berasal dari salah satu aplikasi *E-Wallet* terkemuka di Indonesia, yaitu aplikasi Dana. Data yang diambil merupakan ulasan nyata yang mencerminkan kondisi aktual dan opini pengguna mengenai aplikasi Dana pada saat pengambilan data. Setiap ulasan memberikan wawasan berharga tentang pengalaman pengguna, termasuk kepuasan, masalah yang dihadapi, serta saran untuk perbaikan. Dengan menggunakan data ulasan yang autentik ini, penelitian bertujuan untuk mendapatkan gambaran yang akurat tentang persepsi pengguna terhadap aplikasi Dana, yang kemudian akan dianalisis lebih lanjut untuk mendukung tujuan penelitian.

C. Preprocessing

Preprocessing adalah tahapan awal dalam klasifikasi teks yang bertujuan untuk mempersiapkan data teks sebelum diproses lebih lanjut. Proses ini melibatkan berbagai transformasi data untuk meningkatkan kualitas dan konsistensi teks, seperti penghapusan karakter khusus, normalisasi teks, dan penghilangan kata-kata yang tidak relevan. Dengan melakukan preprocessing, data teks diubah menjadi format yang lebih terstruktur dan bersih, sehingga informasi yang dihasilkan menjadi lebih optimal dan siap digunakan dalam langkah-langkah analisis dan pemodelan berikutnya[6]. Pada penelitian ini metode *Preprocessing* yang digunakan meliputi *Case folding, Tokenizing, Cleaning, Stopword removal, dan Stemming* dan *Lemmatization.* Pada tahap *preprocessing* dilakukan menggunakan *Google Colab*, sebuah platform berbasis *cloud* yang memungkinkan pengolahan data secara efisien dengan memanfaatkan berbagai pustaka dan berbagai alat yang tersedia untuk analisis teks.

Case Folding adalah teknik praproses teks yang sederhana namun sangat efektif, meskipun sering terlupakan. Teknik ini digunakan untuk mengatasi masalah ketidakonsistenan dalam penggunaan huruf besar dan kecil dalam data teks, yang sering kali tidak memiliki struktur yang terorganisir[7].

Tokenisasi atau *Tokenizing* adalah proses mengubah teks atau dokumen menjadi bagian-bagian yang lebih kecil yang disebut token. Token tersebut dapat berupa kata-kata, frasa, tanda baca, atau entitas lainnya. Tujuan utama dari tokenisasi adalah untuk memungkinkan komputer memproses dan memahami teks dengan lebih efisien[8].

Dalam proses pembersihan data atau *Cleaning*, beberapa langkah penting dilakukan untuk memastikan kebersihan dan konsistensi data. Langkah-langkah tersebut meliputi mengisi nilai-nilai yang kosong agar tidak ada data yang hilang atau tidak terwakili, meratakan data yang tidak konsisten untuk memastikan format dan standar yang seragam, serta menangani gangguan data seperti nilai-nilai yang anomali atau tidak relevan sehingga kualitas data tetap terjaga dan analisis selanjutnya dapat dilakukan dengan lebih akurat.[9].

Stopword removal adalah proses dalam pemrosesan teks yang bertujuan untuk menghapus kata-kata yang umum dan sering muncul dalam teks, namun memiliki sedikit atau bahkan tidak ada nilai informatif dalam analisis teks[10]. Kata-kata ini disebut stopwords. Proses stopword removal biasanya dilakukan sebagai salah satu tahapan dalam *preprocessing teks* sebelum analisis lebih lanjut dilakukan[11].

Stemming merupakan langkah dalam pemrosesan teks yang bertujuan untuk menghasilkan bentuk dasar atau kata dasar dari kata-kata yang terdapat dalam teks[12]. Fokus utama dari proses *stemming* adalah untuk menyederhanakan variasi kata dalam teks dengan menghilangkan akhiran kata sehingga dapat diperoleh bentuk dasarnya. Dengan demikian, *stemming* membantu dalam mengonsolidasikan variasi morfologis dari kata-kata yang memiliki akar yang sama. Sebagai contoh, kata-kata seperti "berlari", "berlari", dan "lari" dapat disederhanakan menjadi bentuk dasar "lari" melalui proses *stemming*. Proses ini berkontribusi dalam meningkatkan konsistensi dan mempermudah analisis teks dalam berbagai penerapan Pemrosesan Bahasa Alami (Natural Language Processing/NLP)[13].

Lematisasi atau *Lemmatization* adalah proses pada pemrosesan teks yang bertujuan untuk menghasilkan bentuk kata dasar atau kata lema dari kata-kata yang ada dalam suatu teks[14]. Tujuan utama dari lemmatisasi adalah untuk menyederhanakan kata-kata dalam teks ke bentuk dasarnya dengan memperhatikan konteks dan makna kata tersebut dalam kalimat. Berbeda dengan stemming, lemmatisasi lebih rumit karena mempertimbangkan struktur gramatikal dan konteks kalimat. Sebagai contoh, kata-kata seperti "menjalankan", "menjalankan", dan "lari" dapat disederhanakan menjadi kata dasar seperti "jalan" atau "lari" sesuai dengan makna dan konteks kalimatnya. Lemmatisasi membantu meningkatkan konsistensi dan akurasi analisis teks dalam berbagai penerapan Pemrosesan Bahasa Alami (Natural Language Processing/NLP)[15].

Skenario pengujian

Studi akan mengadopsi tujuh skenario yang berbeda, dimana setiap skenario mengimplementasikan variasi dalam tahap preprocessing. Setiap skenario akan diuji dengan kombinasi dan urutan teknik preprocessing yang beragam.

Tabel 1 *Skenario pengujian*

Skenario	*Preprocessing*
1	*case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi)*
2	*case folding, cleaning, tokenizing, stemming (sastrawi), lemmatization*
3	*case folding, cleaning, tokenizing, stopword removal, lemmatization*
4	*case folding, cleaning, tokenizing, stopword removal*
5	*case folding, cleaning, tokenizing, stemming (sastrawi)*
6	*case folding, cleaning, tokenizing, lemmatization*
7	*case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi), lemmatization*

III. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan bertujuan untuk mengetahui pengaruh dari pemodelan pemrosesan data teks sentimen analisis pada aplikasi *E – Wallet* DANA yang diambil dari salah satu platform yakni *Google Play Store* yang berjumlah 1000 ulasan. Dataset yang diperoleh dari ulasan tersebut kemudian dipilah dan diberi label menjadi dua kategori utama, yaitu ulasan negatif dan ulasan positif. Proses pelabelan dilakukan berdasarkan sistem rating yang terdapat pada aplikasi di Google Play Store. Ulasan dengan rating 3 hingga 5 dikategorikan sebagai ulasan positif, sedangkan ulasan dengan rating 1 dan 2 dikategorikan sebagai ulasan negatif. Pembagian ini bertujuan untuk membedakan sentimen pengguna yang positif terhadap aplikasi dari sentimen yang negatif, sehingga analisis sentimen dapat dilakukan dengan lebih terfokus dan akurat.

Selanjutnya, dataset yang telah diberi label tersebut dibagi menjadi dua bagian untuk keperluan pelatihan dan pengujian model. Sebanyak 400 ulasan digunakan sebagai data train (pelatihan) dan 600 ulasan lainnya digunakan sebagai data test (pengujian). Pembagian ini dirancang untuk memastikan bahwa model yang dikembangkan dapat dilatih dengan sejumlah data yang memadai dan diuji dengan data yang cukup untuk mengevaluasi kinerjanya. Penelitian ini juga menerapkan berbagai skenario pengujian yang berfokus pada tahap preprocessing teks yang berbeda-beda. Tahap preprocessing adalah langkah penting dalam pemrosesan bahasa alami yang mencakup pembersihan dan persiapan data teks sebelum analisis lebih lanjut. Metode preprocessing yang berbeda diterapkan untuk mengetahui pengaruh masing-masing metode terhadap hasil analisis sentimen. Dalam penelitian ini, klasifikasi dilakukan dengan menggunakan metode Support Vector Machine (SVM) empat kernel. Empat kernel diantaranya ada Linear, RBF, Polynomial dan Sigmoid.

A. Hasil pemrosesan data teks

Pada tahap ini dilakukan pengujian dengan menerapkan serangkaian teknik *preprocessing* diantaranya ada *cleaning data, case folding, tokenisasi, stoprowrd removal* dan *stemmer*. Berikut ini adalah hasil dari serangkaian teknik pada *preprocessing* yang telah dilakukan.

Pada Langkah pertama yakni *Cleaning*, Dimana data akan dibersihkan dari hal – hal yang tidak diperlukan atau sulit diproses oleh system seperti tanda baca, symbol, emotikon dan sebagainya.

Tabel 2 *Cleaning*

No	Content	Content_Cleaning
0	Pengembang anjiiiing 😡😡😡 (Vivo Y53), setelah u...	Pengembang anjiiiing Vivo Y53 setelah u...
1	Tahun 2024 ini maaf saya sangat tidak puas den...	Tahun 024 ini maaf saya sangat tidak puas deng...
2	Tolong Ad gak Dana Lite yang lebih simpel, gak...	Tolong Ad gak Dana Lite yang lebih simpel gak...
3	Kenapa saldo saya tbtb terpotong otomatis pada...	Kenapa saldo saya tbtb terpotong otomatis pada...
4	Sekitar 1 thn yg lalu aplikasi sangat bagus, t...	Sekitar 1 thn yg lalu aplikasi sangat bagus t...

Case Folding bertujuan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil, sehingga memastikan konsistensi dan memudahkan proses analisis lebih lanjut.

Tabel 3 *Case folding*

No	Content_Cleaning	Content_Case Folding
0	Pengembang anjiiiing Vivo Y53 setelah u...	pengembang anjiiiing vivo y53 setelah u...
1	Tahun 024 ini maaf saya sangat tidak puas deng...	tahun 024 ini maaf saya sangat tidak puas deng...
2	Tolong Ad gak Dana Lite yang lebih simpel gak...	tolong ad gak dana lite yang lebih simpel gak...
3	Kenapa saldo saya tbtb terpotong otomatis pada...	kenapa saldo saya tbtb terpotong otomatis pada...
4	Sekitar 1 thn yg lalu aplikasi sangat bagus t...	sekitar 1 thn yg lalu aplikasi sangat bagus t...

Dengan tokenisasi, teks dapat dipisahkan menjadi unit-unit yang lebih kecil, memudahkan untuk dilakukan analisis lebih lanjut seperti analisis sentimen, penghitungan frekuensi kata, atau penerapan model pembelajaran mesin dalam pemrosesan bahasa alami

Tabel 4 *Tokenizing*

No	Content_Case Folding	Content_Tokenizing
1	pengembang anjiiiing vivo y53 setelah u...	[pengembang, anjiiiing, vivo, y53, setelah, up...
2	tahun 024 ini maaf saya sangat tidak puas deng...	[tahun, 024, ini, maaf, saya, sangat, tidak, p...
3	tolong ad gak dana lite yang lebih simpel gak...	[tolong, ad, gak, dana, lite, yang, lebih, sim...
4	kenapa saldo saya tbtb terpotong otomatis pada...	[kenapa, saldo, saya, tbtb, terpotong, otomati...
5	sekitar 1 thn yg lalu aplikasi sangat bagus t...	[sekitar, 1, thn, yg, lalu, aplikasi, sangat, ...

Selanjutnya adalah tahap *stopword removal*, yaitu disini menghilangkan kata-kata yang kurang efisien seperti “yang”, “di” dan sebagainya. Dalam bahasa Indonesia, contoh *stopwords* meliputi "yang", "dan", "di", "dari", "itu", dan lain-lain. Dengan menghapus *stopwords* dari teks, kita dapat meningkatkan kualitas analisis teks karena hanya kata-kata yang lebih berarti dan informatif yang tetap dipertahankan. Ini dapat membantu meningkatkan keakuratan analisis sentimen, pengelompokan dokumen, atau tugas-tugas pemrosesan teks lainnya. Tabel dibawah merupakan hasil dari pemfilteran tersebut.

Tabel 5 *Stopword removal*

	content	content_token
0	pengembang anjiiiing vivo y53 setelah u...	[pengembang, anjiiiing, vivo, y53, update, apl...
1	tahun 024 ini maaf saya sangat tidak puas deng...	[024, maaf, puas, aplikasi, dana, skali, tdk, ...
2	tolong ad gak dana lite yang lebih simpel gak...	[tolong, ad, gak, dana, lite, simpel, gak, ber...

Tahap berikutnya adalah *Stemming*. Pada tahap ini akan dilakukan pengubahan beberapa kata menjadi kata dasar sesuai dengan format bahasa Indonesia agar lebih mudah untuk diolah dan berfungsi untuk menigkatkan hasil akurasi.

Tabel 6 *Stemming*

	content	content_token	stemmed
0	pengembang anjiiiing vivo y53 setelah u...	[pengembang, anjiiiing, vivo, y53, update, apl...	[kembang, anjiiiing, vivo, y53, update, aplika...
1	tahun 024 ini maaf saya sangat tidak puas deng...	[024, maaf, puas, aplikasi, dana, skali, tdk, ...	[024, maaf, puas, aplikasi, dana, skali, tdk, ...
2	tolong ad gak dana lite yang lebih simpel gak...	[tolong, ad, gak, dana, lite, simpel, gak, ber...	[tolong, ad, gak, dana, lite, simpel, gak, ber...
3	kenapa saldo saya tbtb terpotong otomatis pada...	[saldo, tbtb, terpotong, otomatis, berlanggana...	[saldo, tbtb, potong, otomatis, langgan, aplik...

Proses lematisasi mempertimbangkan konteks gramatikal kata dalam sebuah kalimat. Misalnya, kata-kata seperti "menyanyikan", "menyanyi", dan "menyanyikan" akan diubah menjadi bentuk dasar "nyanyi". Perbedaan utama antara lematisasi dan stemming adalah bahwa lematisasi menghasilkan kata dasar yang valid secara tata bahasa, sementara stemming hanya memotong akhiran kata tanpa mempertimbangkan konteks.

Tabel 7 *Lemmatization*

	content	content_token	stemmed	text_string	content_lemmatized
0	pengembang anjiiiing vivo y53 setelah u...	[pengembang, anjiiiing, vivo, y53, update, apl...	[kembang, anjiiiing, vivo, y53, update, aplika...	kembang anjiiiing vivo update aplikasi mental ...	kembang anjiiiing vivo y53 update aplikasi men...
1	tahun 024 ini maaf saya sangat tidak puas deng...	[024, maaf, puas, aplikasi, dana, skali, tdk, ...	[024, maaf, puas, aplikasi, dana, skali, tdk, ...	maaf puas aplikasi dana skali kirim uang lapor...	024 maaf puas aplikasi dana skali tdk kirim ua...
2	tolong ad gak dana lite yang lebih simpel gak...	[tolong, ad, gak, dana, lite, simpel, gak, ber...	[tolong, ad, gak, dana, lite, simpel, gak, ber...	tolong dana lite simpel berat ribet iklan cont...	tolong ad gak dana lite simpel gak berat gak r...
3	kenapa saldo saya tbtb terpotong otomatis pada...	[saldo, tbtb, terpotong, otomatis, berlanggana...	[saldo, tbtb, potong, otomatis, langgan, aplik...	saldo tbtb potong otomatis langgan aplikasi me...	saldo tbtb potong otomatis langgan aplikasi me...

B. Analisis hasil pengujian

Pada skenario pertama dilakukan pengujian (*case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi)*) dilanjut (*case folding, cleaning, tokenizing, stemming (sastrawi), lemmatization*), (*case folding, cleaning, tokenizing, stopword removal*), (*case folding, cleaning, tokenizing, stemming (sastrawi)*), (*case folding, cleaning, tokenizing, lemmatization*), (*case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi), lemmatization*). Pada Tabel 8 merupakan hasil akurasi klasifikasi dengan menggunakan metode *Suport Vector Mechine* (SVM) yang telah diuji.

Tabel 8 *Hasil skenario pengujian*

	skenario	akurasi (SVM)				Rata - Rata
		Linear	RBF	Polynomial	Sigmoid	
1	case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi)	100%	100%	99,42%	94,25%	98,42%
2	case folding, cleaning, tokenizing, stemming (sastrawi), lemmatization	100%	100%	99,42%	98,85%	99,57%
3	case folding, cleaning, tokenizing, stopword removal, lemmatization	100%	100%	98%	96,50%	98,63%
4	case folding, cleaning, tokenizing, stopword removal	100%	100%	99,42%	97,12%	99,14%
5	case folding, cleaning, tokenizing, stemming (sastrawi)	100%	100%	99,42%	98,27%	99,42%
6	case folding, cleaning, tokenizing, lemmatization	100%	100%	99%	99,50%	99,63%
7	case folding, cleaning, tokenizing, stopword removal, stemming (sastrawi), lemmatization	98,08%	98,08%	94,25%	93,30%	95,93%

Tabel 8 menyajikan hasil pengujian pada dataset ulasan aplikasi Dana yang menggunakan berbagai teknik preprocessing dan empat kernel berbeda dari Support Vector Machine (SVM). Hasil menunjukkan perbedaan signifikan dalam akurasi berdasarkan kombinasi teknik preprocessing dan jenis kernel yang diterapkan.

Pada skenario pertama, dengan teknik *case folding, cleaning, tokenizing, stopword removal*, dan *stemming* menggunakan algoritma Sastrawi, semua kernel mencapai akurasi 100% kecuali kernel Polynomial (99,42%) dan Sigmoid (94,25%) dengan rata – rata (98,42%). Skenario kedua menambahkan *lemmatization* ke dalam teknik yang sama, menghasilkan akurasi yang identik dengan skenario pertama untuk kernel Linear dan RBF, tetapi dengan penurunan sedikit pada kernel Polynomial (99,42%) dan peningkatan pada kernel Sigmoid (98,85%) yang menghasilkan rata – rata (99,57%). Pada skenario ketiga, dengan mengganti *stemming* dengan *lemmatization* setelah menghapus *stopwords*, kernel Linear dan RBF masih mencapai akurasi 100%, tetapi kernel Polynomial dan Sigmoid menunjukkan penurunan akurasi menjadi 98% dan 96,50% masing-masing dengan rata – rata (98,63%). Skenario keempat, yang hanya mencakup case *folding, cleaning, tokenizing, dan stopword removal*, menghasilkan akurasi yang konsisten tinggi pada kernel Linear dan RBF (100%), dengan kernel Polynomial sedikit menurun menjadi 99,42% dan kernel Sigmoid meningkat menjadi 97,12% dan rata – rata (99,14%). Pada skenario kelima, dengan *case folding, cleaning, tokenizing*, dan *stemming* tanpa menghapus *stopwords*, akurasi kernel Linear dan RBF tetap 100%, sementara kernel Polynomial mencatat 99,42% dan kernel Sigmoid meningkat menjadi 98,27% dan rata – rata (99,42%). Skenario keenam, yang menggunakan *case folding, cleaning, tokenizing*, dan *lemmatization* tanpa menghapus *stopwords*, menunjukkan akurasi terbaik di antara semua skenario, dengan kernel Linear dan RBF tetap 100%, kernel Polynomial sedikit menurun menjadi 99%, dan kernel Sigmoid mencapai 99,50% dengan rata – rata (99,63%). Terakhir, skenario ketujuh, yang menggabungkan *case folding, cleaning, tokenizing, stopword removal, stemming*, dan *lemmatization*, menunjukkan penurunan akurasi dengan kernel Linear dan RBF masing-masing mencapai 98,1%, kernel Polynomial 94,25%, dan kernel Sigmoid 93,30% dan rata – rata (95,93%).

Secara keseluruhan, teknik preprocessing yang menggabungkan lemmatization memberikan hasil terbaik dalam hal akurasi, terutama dengan kernel Linear dan RBF yang menunjukkan konsistensi tinggi. Kernel Polynomial dan Sigmoid memperlihatkan variasi akurasi yang lebih besar tergantung pada teknik preprocessing yang digunakan, dengan skenario keenam menonjol sebagai hasil terbaik di antara semua pengujian dengan hasil rata – rata 99,63%.

IV. SIMPULAN

Berdasarkan hasil dari beberapa skenario pengujian yang telah dilakukan, pengaruh preprocessing terhadap berbagai model teknik ditentukan untuk menghasilkan tingkat akurasi yang lebih tinggi. Dari penelitian tersebut dapat disimpulkan bahwa system kinerja terbaik dihasilkan ketika menggunakan kombinasi *case folding, cleaning, tokenizing, dan lemmatization* yang menghasilkan tertinggi diantara teknik lain. Skenario pengujian diuji menggunakan metode SVM menggunakan keempat kernel yang menghasilkan 100% pada kernel *Linear,* 100% pada kernel *RBF,* 99% pada kernel *Polynomial*, 99.50% pada kernel *Sigmoid.* Peneliti menemukan bahwa dengan menggunakan seluruh teknik *preprocessing* pada skenario pengujian terakhir belum tentu meningkatkan keakurasian.

Untuk penelitian selanjutnya, bisa menambahkan jumlah dataset yang digunakan serta menguji berbagai teknik preprocessing tambahan untuk meningkatkan tingkat akurasi dan mengurangi kesalahan klasifikasi. Beberapa teknik preprocessing yang dapat dieksplorasi termasuk normalisasi, standardisasi, penghapusan data outlier, pengisian data yang hilang, dan teknik reduksi dimensi seperti Principal Component Analysis (PCA). Selain itu, penting juga untuk mencoba teknik ekstraksi fitur yang lebih canggih seperti metode berbasis wavelet atau transformasi Fourier. Penelitian juga dapat mempertimbangkan untuk menggunakan teknik augmentasi data untuk memperluas dataset secara artifisial, yang dapat membantu dalam meningkatkan performa model. Evaluasi pengaruh setiap teknik preprocessing terhadap kinerja model secara individual dan dalam kombinasi yang berbeda juga dapat memberikan wawasan berharga tentang metode mana yang paling efektif.

UCAPAN TERIMA KASIH

Peneliti ingin menyampaikan rasa terima kasih yang mendalam kepada semua pihak yang telah mendukung peneliti dalam perjalanan ini. Terima kasih kepada kepada Kaprodi dan seluruh staf Program Studi Teknik Informatika di Universitas Muhammadiyah Sidoarjo (UMSIDA) atas dukungan dan bimbingannya. Terutama, Peneliti mengucapkan terima kasih kepada dosen pembimbing Peneliti atas bimbingan, arahan, dan kesabaran luar biasa yang telah beliau berikan selama penelitian ini. Dukungan beliau sangat membantu peneliti dalam melewati setiap tahapan penelitian. Terimakasih kepada keluarga Peneliti atas doa, dukungan, dan kesabaran mereka selama proses ini. Ucapan terima kasih juga peneliti tujukan kepada teman-teman dan rekan-rekan sejawat yang telah memberikan dukungan moral dan semangat yang tak terhingga. Selain itu, peneliti berterima kasih Semoga hasil penelitian ini dapat memberikan kontribusi yang berarti bagi kemajuan ilmu pengetahuan. Terima kasih atas segala dukungan yang telah diberikan.

REFERENSI

[1] N. D. Abrilia and S. Tri, "Pengaruh Persepsi Kemudahan Dan Fitur Layanan Terhadap Minat Menggunakan E-Wallet Pada Aplikasi Dana Di Surabaya," *J. Pendidik. Tata Niaga*, vol. 8, no. 3, pp. 1006–1012, 2020.

[2] S. P. Anggraini and S. Suaidah, "Sistem Informasi Sentral Pelayanan Publik dan Administrasi Kependudukan Terpadu dalam Peningkatan Kualitas Pelayanan Kepada Masyarakat Berbasis Website …," *J. Teknol. Dan Sist. Inf.*, vol. 3, no. 1, pp. 12–19, 2022, [Online]. Available: http://jim.teknokrat.ac.id/index.php/sisteminformasi/article/view/1658%0Ahttp://jim.teknokrat.ac.id/index.php/sisteminformasi/article/viewFile/1658/579

[3] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1120–1126, Dec. 2021, doi: 10.29207/resti.v5i6.3588.

[4] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[5] R. Ulgasesa, A. B. P. Negara, and T. Tursina, "Pengaruh Stemming Terhadap Performa Klasifikasi Sentimen Masyarakat Tentang Kebijakan New Normal," *J. Sist. dan Teknol. Inf.*, vol. 10, no. 3, p. 286, 2022, doi: 10.26418/justin.v10i3.53880.

[6] Z. R. N. S. Prasetija, A. Romadhony, and E. B. Setiawan, "Analisis Pengaruh Normalisasi Teks pada Klasifikasi Sentimen Ulasan Produk Kecantikan," *e-Proceeding Eng.*, vol. 9, no. 3, pp. 1769–1775, 2022, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18184/17795

[7] H. H. Mubaroroh, H. Yasin, and A. Rusgiyono, "Analisis Sentimen Data Ulasan Aplikasi Ruangguru Pada Situs Google Play Menggunakan Algoritma Naïve Bayes Classifier Dengan Normalisasi Kata Levenshtein Distance," *J. Gaussian*, vol. 11, no. 2, pp. 248–257, 2022, doi: 10.14710/j.gauss.v11i2.35472.

[8] G. A. BUNTORO, R. ARIFIN, G. N. SYAIFUDDIIN, A. SELAMAT, O. KREJCAR, and H. FUJITA,

"Implementation of a Machine Learning Algorithm for Sentiment Analysis of Indonesia's 2019 Presidential Election," *IIUM Eng. J.*, vol. 22, no. 1, pp. 78–92, 2021, doi: 10.31436/IIUMEJ.V22I1.1532.
[9] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.
[10] A. B. Putra Negara, "The Influence Of Applying Stopword Removal And Smote On Indonesian Sentiment Classification," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 14, no. 3, p. 172, 2023, doi: 10.24843/lkjiti.2023.v14.i03.p05.
[11] S. J. Angelina, A. Bijaksana, P. Negara, and H. Muhardi, "Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia," vol. 02, no. 1, pp. 165–173, 2023, doi: 10.26418/juara.v2i1.69680.
[12] H. A. Almuzaini and A. M. Azmi, "Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020, doi: 10.1109/ACCESS.2020.3009217.
[13] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based Dan Random Forest," *J. Ilm. Inform.*, vol. 11, no. 02, pp. 159–169, 2023, doi: 10.33884/jif.v11i02.7987.
[14] F. Noer Azzahra *et al.*, "Penerapan Metode Naive Bayes Dalam Klasifikasi Spam SMS Menggunakan Fitur Teks Untuk Mengatasi Ancaman Pada Pengguna," *J. Inf. Syst. Res.*, vol. 5, no. 3, p. 880, 2024, doi: 10.47065/josh.v5i3.5070.
[15] D. A. Vonega, A. Fadila, and D. E. Kurniawan, "Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024," *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 129–135, 2022, doi: 10.30871/jaic.v6i2.4300.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.