ANALISIS SENTIMEN PADA ULASAN PRODUK *CETAPHIL GENTLE SKIN CLEANSER* DI WEBSITE *FEMALE DAILY* MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* (SVM)

SKRIPSI

Oleh:
CITRA KHAERUN NISA
NIM. 200605110132

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

ANALISIS SENTIMEN PADA ULASAN PRODUK *CETAPHIL GENTLE SKIN CLEANSER* DI WEBSITE FEMALE DAILY MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* (SVM)

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan dalam Memperoleh
Gelar Sarjana Komputer (S.Kom)

Oleh:
CITRA KHAERUN NISA
NIM. 200605110132

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

HALAMAN PERSETUJUAN

ANALISIS SENTIMEN PADA ULASAN PRODUK *CETAPHIL GENTLE SKIN CLEANSER* DI WEBSITE FEMALE DAILY MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* (SVM)

SKRIPSI

Oleh:
CITRA KHAERUN NISA
NIM. 200605110132

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 6 Juni 2024

Pembimbing I,	Pembibing II,
Dr. Totok Chamidy, M.Kom	Prof. Dr.Suhartono, M.Kom
NIP. 19691222 200604 1 001	NIP. 19680519 200312 1 001

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

ANALISIS SENTIMEN PADA ULASAN PRODUK *CETAPHIL GENTLE SKIN CLEANSER* DI WEBSITE FEMALE DAILY MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* (SVM)

SKRIPSI

Oleh:
CITRA KHAERUN NISA
NIM. 200605110132

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal:14 Juni 2024

Susunan Dewan Penguji

Ketua Penguji	: Dr. Zainal Abidin, M.Kom NIP. 19760613 200501 1 004	( )
Anggota Penguji I	: Dr. M. Ainul Yaqin, M.Kom NIP. 19761013 200604 1 004	( )
Anggota Penguji II	: Dr. Totok Chamidy, M.Kom NIP. 196912229 200604 1 001	( )
Anggota Penguji III	: Prof. Dr. Suhartono, S.Si., M.Kom NIP. 19680519 200312 1 001	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Citra Khaerun Nisa
NIM	: 200605110132
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Analisis Sentimen Pada Ulasan Produk Cetaphil Gentle Skin Cleanser Di Website Female Daily Menggunakan Metode *Support Vector Machine* (SVM)

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 14 Juni 2024
Yang membuat pernyataan,

Citra Khaerun Nisa
NIM. 200605110132

HALAMAN MOTTO

“If it is meant for you, it will find you.”

“Mulai aja dulu!”

HALAMAN PERSEMBAHAN

Saya persembahkan karya ini kepada:

Ayah saya,

Kaderi

Yang telah mendukung dan menyemangati saya hingga sampai pada titik ini

Ibu saya,

Jauhariah

Yang telah mendukung dan menyemangati saya hingga sampai pada titik ini

Kakak-kakak dan adik saya,

Harieza Citra R.J, Anenda Citra Rizqi, dan Daffa Raqilla Al Farid

Yang telah mendukung dan menyemangati saya hingga sampai pada titik ini

Sahabat-sahabat semasa sekolah saya,

Aulia Zulfaeda, Devi Kamalia Safitri, Hernawati, Rohyatul Audil, Farida Fasa, dan Natasha Agustina

Yang telah bersedia mendengarkan keluh kesah saya dan menyemangati saya hingga sampai pada titik ini

Teman-teman seperjuangan,

Teknik Informatika Angkatan 2020

Semoga kita semua selalu diberi kemudahan oleh Allah SWT

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

Segala puji hanya milik Allah Subhanahu Wa Ta'ala atas segala nikmat dan kasih sayang-Nya yang telah memudahkan penulis untuk menyelesaikan skripsi dengan judul "Analisis Sentimen Pada Ulasan Produk Cetaphil Gentle Skin Cleanser Di Website Female Daily Menggunakan Metode Support Vector Machine (SVM)". Semoga shalawat dan salam senantiasa terlimpah kepada Nabi Muhammad Sallalahu 'Alaihi wa Sallam. Semoga kita semua mendapat syafaatnya di hari akhir kelak, Aamiin.

Penulis mengucapkan rasa syukur dan terima kasih yang tak terhingga kepada semua pihak-pihak yang selalu memberikan bantuan dan motivasi kepada penulis sehingga dapat menyelesaikan skripsi ini. Ucapan ini penulis sampaikan kepada:

1. Prof. Dr. H. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, M.MT., IPM, selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Totok Chamidy, M.Kom selaku dosen pembimbing I dan Prof. Dr. Suhartono, S.Si., M.Kom selaku dosen pembimbing II yang telah memberikan bantuan dan arahan kepada penulis, sehingga bisa menyelesaikan skripsi ini.

5. Dr. Zainal Abidin, M.Kom selaku dosen penguji I dan Dr. M. Ainul Yaqin, M.Kom selaku dosen penguji II yang telah menguji serta memberikan masukan sehingga penulis dapat menyelesaikan skripsi dengan baik.
6. Segenap Dosen, Admin, Laboran dan Mahasiswa Program Studi Teknik Informatika yang telah memberikan banyak dukungan, bantuan, dan bimbingan selama perkuliahan.
7. Orang tua serta kedua kakak dan adik penulis yang selalu memberikan semangat untuk terus berusaha serta doa yang tak putus-putusnya selalu dicurahkan untuk setiap proses yang penulis lalui sehingga dapat menyelesaikan skripsi ini.
8. Sahabat-sahabat saya Aulia Zulfaeda, Devi Kamalia Safitri, Hernawati, Rohyatul Audil, Farida Fasa, dan Natasha Agustina yang telah bersedia menyisihkan waktunya untuk mendengar berbagai keluh kesah dan cerita penulis selama menjadi mahasiswa rantau dan selalu menyamangati penulis sehingga dapat menyelesaikan skripsi ini.
9. Teman satu program studi sekaligus teman PKL dan teman kos penulis Khalimatul Luthfiyyah yang telah menemani penulis selama masa perkuliahan.
10. Lee Haechan, Na Jaemin, dan member NCT lainnya yang secara tidak langsung memberikan warna dan motivasi untuk penulis melalui karya-karyanya.

Akhir kata, penulis mengakui bahwa penulisan pada skripsi ini masih banyak kekurangan. Saya berharap semoga skripsi ini diterima sebagai amal ibadah yang tulus dan bermanfaat di sisi Allah Subhanahu Wa Ta'ala. Semoga karya ini menjadi

bagian dari kontribusi yang tak terputus dalam rangka memperkuat dan mengembangkan ilmu pengetahuan, serta melaksanakan tugas sebagai hamba Allah yang berkomitmen.

Wassalamualaikum Warahmatullahi Wabarakatuh.

Malang, 14 Juni 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Nisa, Citra Khaerun. 2024. **Analisis Sentimen Pada Ulasan Produk Cetaphil Gentle Skin Cleanser Di Website Female Daily Menggunakan Metode Support Vector Machine (SVM).** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Totok Chamidy, M.Kom. (II) Prof. Dr. Suhartono, S.Si., M.Kom

Kata kunci: *Analisis Sentimen, Cetaphil, Female Daily, Support Vector Machine.*

Industri kecantikan kini telah menjadi salah satu sektor berkembang pesat di Indonesia. Banyaknya produk kecantikan yang ada membuat konsumen semakin selektif dalam memilih suatu produk, khususnya produk berupa *facial wash. Facial wash* merupakan salah satu produk *skincare* yang sering dicari oleh konsumen. Banyaknya *brand* yang menyediakan *facial wash* seringkali membuat konsumen semakin bingung dalam memilih produk yang sesuai dengan preferensi kulit mereka. Oleh karena itu, *review* sebuah produk dari pengguna sebelumnya akan menjadi informasi yang sangat berharga. Ulasan-ulasan pengguna menjadi topic menarik untuk diteliti, karena ada berbagai penilaian dan sentimen terkait produk tersebut. Tujuan dari penelitian ini adalah utuk menganalisis sentimen pengguna produk Cetaphil Gentle Skin Cleanser pada website *Female Daily* menggunakan *Support Vector Machine.* Dalam penelitian ini diperoleh data sebanyak 1050 data yang terdiri dari 688 data berlabel positif dan 362 data berlabel negatif. Dari penelitian ini diketahui bahwa algoritma Support Vector Machine dapat bekerja dengan cukup baik pada analisis sentimen. Nilai akurasi terbaik didapat melalui hasil uji coba dengan rasio 9:1 pada proporsi positif negatif 5:5 saat fungsi kernel Linear menggunakan nilai *hyperparameter* C = 1. Uji coba tersebut menghasilkan nilai akurasi sebesar 87%. Selain itu, melalui uji coba lain diperoleh hasil bahwa terdapat beberapa aspek yang mempengaruhi performa sistem seperi kombinasi proses *preprocessing,* rasio pembagian data latih dan uji, proporsi data berlabel positif dan negatif pada data latih, serta pemilihan fungsi kernel.

ABSTRACT

Nisa, Citra Khaerun. 2024. **Sentiment Analysis on Cetaphil Gentle Skin Cleanser Product Reviews on the Female Daily Website Using the Support Vector Machine (SVM)** Method. Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Advisor: (I) Dr. Totok Chamidy, M.Kom. (II) Prof. Dr. Suhartono, S.Si., M.Kom.

The beauty industry has witnessed exponential growth in Indonesia. The proliferation of beauty products has compelled consumers to exercise greater selectivity in their product choices, particularly in the case of facial wash. Facial wash is a highly sought-after skincare product. The plethora of brands offering facial wash has made it challenging for consumers to identify a product that aligns with their skin preferences. In such a scenario, user reviews can serve as invaluable information. The topic of user reviews is worthy of further investigation, as they offer a wealth of information about the product in question, including a variety of assessments and sentiments. The objective of this research is to analyze the sentiment of Cetaphil Gentle Skin Cleanser product users on the Female Daily website using Support Vector Machine (SVM). In this study, 1050 data points were obtained, comprising 688 positively labeled data points and 362 negatively labeled data points. The results demonstrate that the SVM algorithm can effectively perform sentiment analysis. The optimal accuracy value was achieved through experimental results with a ratio of 9:1 at a positive-negative proportion of 5:5 when the Linear kernel function utilized a hyperparameter value of C = 1. The experimental trial yielded an accuracy value of 87%. Furthermore, additional trials revealed that several factors influence the system's performance, including the combination of preprocessing techniques, the ratio of training and test data, the proportion of positive and negative labeled data in the training data, and the selection of kernel functions.

Keywords: *Sentiment Analysis, Cetaphil, Female Daily, Support Vector Machine.*

مستخلص البحث

نيسا، سترة خيرون. 2024. تحليل المشاعر على مراجعات منتج منظف البشرة اللطيف من سيتافيل على الموقع الإلكتروني اليومي للإناث باستخدام طريقة آلة دعم المتجهات .(SVM) الأطروحة. برنامج دراسة هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف: (الأول) الدكتور توتوك شاميدي، ماجستير كوم (II) .البروفيسور الدكتور سوهارتونو، س. س، م. كوم.

أصبحت صناعة التجميل الآن واحدة من أسرع القطاعات نموًا في إندونيسيا. إن عدد منتجات التجميل الموجودة تجعل المستهلكين أكثر انتقائية في اختيار المنتج، وخاصة المنتجات التي على شكل غسول الوجه. غسول الوجه هو أحد منتجات العناية بالبشرة التي غالبًا ما يبحث عنها المستهلكون. غالبًا ما يجعل عدد العلامات التجارية التي توفر غسول الوجه المستهلكين أكثر حيرة في اختيار المنتج الذي يناسب تفضيلات بشرتهم. ولذلك، فإن مراجعات المستخدمين السابقين للمنتج ستكون معلومات قيمة للغاية. تُعد مراجعات المستخدمين موضوعًا مثيرًا للاهتمام للبحث، حيث أن هناك العديد من الأحكام والمشاعر المتعلقة بالمنتج . يتمثل الغرض من هذا البحث في تحليل مشاعر مستخدمي منظف البشرة اللطيف من سيتافيل على الموقع الإلكتروني لـ "فاميلي ديلي" باستخدام آلة ناقلات الدعم. في هذه الدراسة، تم الحصول على 1050 بيانات تتألف من 688 بيانات ذات تقييم إيجابي و362 بيانات ذات تقييم سلبي. من هذا البحث، من المعروف أن خوارزمية آلة ناقلات الدعم يمكن أن تعمل بشكل جيد في تحليل المشاعر. تم الحصول على أفضل قيمة دقة من خلال النتائج التجريبية بنسبة 9:1 مع نسبة إيجابية إلى سلبية بنسبة 5:5 عندما تستخدم دالة النواة الخطية قيمة المعرف الفائق = C 1 .بالإضافة إلى ذلك، ومن خلال تجارب أخرى، يتبين أن هناك العديد من الجوانب التي تؤثر على أداء النظام مثل الجمع بين عمليات المعالجة المسبقة، ونسبة بيانات التدريب وبيانات الاختبار، ونسبة البيانات الموسومة إيجابًا وسلبًا في بيانات التدريب. واختيار دالة النواة.

الكلمات المفتاحية: تحليل المشاعر، سيتافيل، أنثى يومي، آلة دعم المتجهات

BAB I

PENDAHULUAN

1.1 Latar Belakang

Industri kecantikan kini telah menjadi salah satu sektor berkembang pesat di Indonesia. Perkembangan ini dapat dilihat melalui pertumbuhan jumlah industri kecantikan yang mencapai 1.010 perusahaan pada pertengahan tahun 2023. Kesadaran konsumen akan penampilan yang sehat dan perawatan diri inilah yang meningkatkan permintaan terhadap produk kecantikan atau perawatan kulit. Banyaknya produk kecantikan yang ada membuat konsumen semakin selektif dalam memilih suatu produk, khususnya produk berupa *facial wash. Facial Wash* merupakan salah satu produk *skincare* yang sering dicari oleh konsumen. Dalam memilih *facial wash,* biasanya konsumen benar-benar memperhatikan kualitas *facial wash* karena sangat berperan penting untuk merawat kulit wajah. Pemilihan *facial wash* yang salah dapat menyebabkan iritasi kulit. Banyaknya *brand* yang menyediakan *facial wash* seringkali membuat konsumen semakin bingung dalam memilih produk yang sesuai dengan preferensi kulit mereka. Oleh karena itu, *review* sebuah produk dari pengguna sebelumnya akan menjadi informasi yang sangat berharga. Seiring dengan perkembangan internet dan teknologi, *review* produk dapat dengan mudah ditemukan.

Internet memiliki dampak besar terhadap cara individu berperilaku dan berinteraksi beberapa tahun terakhir. Media sosial menjadi beberapa contoh dampak perkembangan teknologi tersebut. Media sosial`kini dijadikan sebagai sarana utama untuk berkomunikasi, mencari hiburan, hingga sebagai saluran

berbagi berita dan berbagi informasi terkini. Selain itu, media sosial juga telah menjadi platform yang memungkinkan individu dapat mengekspresikan diri, berkreasi, menemukan komunitas dengan minat yang sama, hingga untuk memperluas relasi sosial (Martini & Dewi, 2021).

Kemunculan media sosial banyak didukung oleh kebutuhan serta tuntutan dari *user* itu sendiri, salah satunya contohnya adalah Female Daily. Female Daily merupakan salah satu forum yang menjadi wadah bertukar informasi terkait produk kecantikan seperti *make up* dan *skincare* termasuk *facial wash*. Pada tahun 2023, aplikasi Female Daily telah diunduh kurang lebih sebanyak 1.172.000 kali berdasarkan data yang tersedia di websitenya. Dalam forum ini, pengguna dapat dengan bebas memberikan penilaian terhadap produk tanpa dibatasi karakter sehingga kualitas produk dapat diketahui secara detail.

Ulasan atau *review* produk yang diposting pengguna tentu saja dapat memberikan wawasan yang sangat berguna terkait kepuasan dan pengalaman pengguna terhadap suatu produk. Membaca ulasan produk biasanya dapat membuat calon konsumen lebih tertarik pada produk yang akan dibeli. Ulasan produk yang diberikan pengguna menggambarkan keuntungan atau kerugian yang akan didapatkan apabila sebuah produk digunakan. Ulasan produk bernilai positif otomatis akan memunculkan persepsi positif bagi suatu *brand* kecantikan dan *skincare* seperti Cetaphil. Pada penelitian ini akan dilakukan analisis sentimen untuk mengetahui tentang kecenderungan opini pengguna terhadap ulasan untuk produk *Facial Wash* Cetaphil.

Cetaphil merupakan salah satu *brand* perawatan kulit yang telah berdiri sejak tahun 1947 dan banyak direkomendasikan oleh dokter. Meskipun telah lama berdiri, eksistensi *brand* ini tidak meredup sama sekali. Cetaphil Gentle Skin Cleanser menjadi salah satu produk yang banyak dikenal oleh masyarakat saat ini. *Facial Wash* ini telah terjual 10 ribu produk di salah satu e-commerce. Banyaknya penjualan produk Cetaphil Gentle Skin Cleanser ini tentu saja menyebabkan jumlah ulasan di sosial media termasuk Female Daily semakin banyak juga. Ulasan produk yang diberikan dapat berpengaruh kepada citra *brand* serta jumlah penjualan kedepannya. Maka untuk untuk mempertahankan hal tersebut, perlu dilakukan analisis sentimen terkait ulasan yang diberikan kemudian diproses menggunakan *text mining.*

Analisis sentimen atau opinion mining yang termasuk salah satu bidang dalam *Naturan Language Processing* (NLP) adalah proses pemahaman, penggalian, serta pengolahan data tekstual secara otomatis untuk mendapat informasi sentimen yang terkandung dalam suatu opini (Karim, 2020). Analisis Sentimen dilakukan untuk melihat opini seseorang yang ditujukan ke berbagai hal (Buntoro, 2017). Biasanya analisis sentimen akan membuat kategori dalam opini yang dianalisis, yakni label sentimen positif, netral, atau negatif. Dokumen, berupa teks opini, kata-kata dan sebagainya semuanya dapat diklasifikasikan menurut tingkat polaritasnya menggunakan analisis sentimen.. Klasifikasi ini bertujuan untuk memahami apakah teks tersebut termasuk ke dalam label yang ditentukan. Maksud dari polaritas adalah sebuah kecenderungan yang menentukan suatu teks berupa opini, dokumen, atau pernyataan memiliki sifat positif atau negatif.

Text Mining adalah pencarian data-data berbentuk teks yang sebelumnya tidak diketahui, sehingga nantinya dapat dimengerti (Firdaus & Firdaus, 2021). Text Mining merupakan salah satu teknik yang digunakan dalam melakukan klasifikasi dokumen (Herianto, 2019). Adapun Text Mining menjadi salah satu variasi dari Data Mining yang nantinya akan menemukan pola menarik dari sekumpulan data tekstual yang berjumlah besar. Text mining memiliki beberapa bagian, salah satunya adalah analisis sentimen (Rochmawati & Wibawa, 2018).

Permasalah terkait analisis sentimen sangat perlu untuk diperhatikan, khususnya bagi seorang muslim untuk dapat memperjelas opini. Seringkali masyarakat dalam beropini tidak memperhatikan tutur kata yang baik. Selain itu, banyak pengguna sosial media yang sering mencari-cari kesalahan terhadap sesuatu, entah itu produk ataupun pihak tertentu yang dirasa terlibat dalam suatu hal. Padahal Allah SWT telah berfirman dalam Al-Qur'an surat Al-Hujurat ayat 12 yang berbunyi:

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُواْ ٱجْتَنِبُواْ كَثِيرًا مِّنَ ٱلظَّنِّ إِنَّ بَعْضَ ٱلظَّنِّ إِثْمٌۖ وَلَا تَجَسَّسُواْ وَلَا يَغْتَب بَّعْضُكُم بَعْضًاۚ أَيُحِبُّ أَحَدُكُمْ أَن يَأْكُلَ لَحْمَ أَخِيهِ مَيْتًا فَكَرِهْتُمُوهُۚ وَٱتَّقُواْ ٱللَّهَۚ إِنَّ ٱللَّهَ تَوَّابٌ رَّحِيمٌ

"Hai orang-orang yang beriman, jauhilah kebanyakan purba-sangka (kecurigaan), karena sebagian dari purba-sangka itu dosa. Dan janganlah mencari-cari keburukan orang dan janganlah menggunjingkan satu sama lain. Adakah seorang diantara kamu yang suka memakan daging saudaranya yang sudah mati? Maka tentulah kamu merasa jijik kepadanya. Dan bertakwalah kepada Allah. Sesungguhnya Allah Maha Penerima Taubat lagi Maha Penyayang." (Q.S Al-Hujurat:12).

Ayat di atas mengajarkan umat Islam untuk menghindari prasangka buruk dan melarang dalam mencari-cari kesalahan orang lain. Prisnsip-prinsip ini sangat relevan dalam konteks analisis sentimen pada era modern seperti sekarang. Analisis

sentimen merupakan pendekatan teknologi yang dilakukan untuk memahami serta mengevaluasi opini, perasaan, ataupun pandangan seseorang melalui teks. Penilaian dan prasangka buruk terhadap sesuatu, orang lain atau kelompok tertentu seringkali menjadi awal mula untuk komentar yang merugikan dalam konteks sosial media. Ketika menerapkan analisis sentimen, prinsip-prinsip tersebut dapat diartikan sebagai pemahaman bahwa mencari-cari kesalahan dan mengekspresikan prasangka buruk terhadap sesuatu melalui media sosial dapat berdampak negatif pada lingkungan media sosial itu sendiri. .

Penelitian terkait analisis sentimen telah banyak dilakukan, salah satunya oleh (Pratiwi *et al*., 2021) yang menganalisis produk *skincare* pada website Female Daily menggunakan *Support Vector Machine* (SVM) dimana *preprocessing* yang diterapkan hanya 3 tahap. Nilai akurasi sebesar 87% diperoleh di penelitian ini. Penelitian lainnya dilakukan oleh (Hamka *et al*., 2022) yang menganalis produk serupa menggunakan *Naïve Bayes Classifier*. Penelitian ini meneliti objek produk kecantikan jenis serum yang datanya diambil dari *Twitter* (sekarang dikenal X) dengan metode *crawling* data. Hasil akhir dari analisis sentimen yang dilakukan menghasilkan nilai akurasi tertinggi sebesar 80%. Penelitian lain terkait analisis sentimen juga dilakukan oleh (Muktafin *et al*., 2020) yang meneliti ulasan pembelian produk pada *marketplace* Shopee dengan pendekatan *Natural Language Processing* (NLP). Melalui enelitian ini, dihasilkan akurasi sebesar 76,92%. Banyaknya metode serta kombinasi yang dapat digunakan dalam analisis sentimen seperti penelitian sebelumnya, maka peneliti kali ini memilih metode SVM untuk melakukan analisis sentimen.

Metode SVM adalah salah satu teknik yang dapat memprediksi kelas berdasar pada pola yang ditemukan melalui hasil proses pelatihan. Metode SVM adalah salah satu algoritma efektif untuk melakukan klasifikasi pada data teks. Prinsip inti dari metode SVM sendiri adalah untuk mengidentifikasi *hyperlane* terbaik untuk membagi dua kelas dalam satu ruang fitur. Meskipun terbilang baru daripada teknik lainnya, teknik ini dapat bekerja lebih baik dalam berbagai bidang yang berhubungan dengan teknologi, misalnya klasifikasi teks, bioinformatika, hingga pengenalan tulisan tangan dan yang lainnya. Dari banyaknya penelitian menggunakan metode SVM, hasil akurasi penggunaan metode SVM cenderung tinggi. Hal inilah yang mendorong penulis memilih untuk menggunakan SVM sebagai metode pada penelitian ini. Melakukan klasifikasi terhadap sentiment positif dan negatif pada ulasan produk yang diberikan pengguna di laman Female Daily adalah tujuan utama dalam penelitian ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, bagaimana efisiensi analisis penggunaan metode *Support Vector Machine* pada analisis sentimen untuk ulasan produk Cetaphil Gentle Skin Cleanser di website Female Daily?

1.3 Batasan Masalah

Untuk menghindari deviasi topik dari penelitian ini, diterapkan beberapa batasan masalah seperti berikut:

1. Data yang digunakan dalam penelitian yaitu data primer berupa teks ulasan dari website Female Daily.

2. Sebanyak 1050 data komentar ulasan pengguna Cetaphil Gentle Skin Cleanser diimplementasikan.
3. Data yang dijadikan bahan analisis adalah ulasan pada website Female Daily tahun 2018 sampai dengan 11 Mei 2024.

1.4 Tujuan Penelitian

Melakukan analisis sentimen pada *review* atau ulasan produk Cetaphil Gentle Skin Cleanse yang diberikan pengguna di laman Female Daily.

1.5 Manfaat Penelitian

Melalui penelitian ini, diharapkan dapat diambil beberapa manfaat seperti berikut:

1. Menambah wawasan serta pengetahuan tentang analisis sentimen, terutama pada media sosial
2. Dapat menjadi acuan untuk penelitian selanjutnya dengan masalah serupa secara lebih mendalam.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terdahulu

Penelitian yang dilakukan oleh Zidna dan kawan-kawan pada tahun 2021 menggunakan *Support Vector Machine* guna menganalisis sentimen pada data tweet terkait dengan marketplace Bukalapak (Alhaq *et al*., 2021). Metode SVM dipilih karena memiliki akurasi yang lebih tinggi dibandingkan dengan metode lain seperti *Naïve Bayes Classifier*. Proses penelitian pada penelitian ini meliputi pengumpulan data tweet dari Twitter menggunakan Twitter Scraper, pembagian data *training* dan data prediksi, pelabelan sentimen secara manual, dan preprocessing data seperti cleaning, tokenizing, dan normalisasi kata. Hasil klasifikasi sentimen dengan SVM mencapai akurasi tertinggi sebesar 93% menggunakan *K-Fold Cross Validation*. Berdasarkan hasilnya, dapat dilihat bahwa penggunaan SVM pada penelitian ini mampu mengklasifikasikan sentimen data tweet dengan tingkat akurasi yang tinggi.

Akbar Zaeim Pragahakusma dan Novrido Charibaldi dalam penelitian yang dilakukan tahun 2021 membandingkan fungsi kinerja dari 3 kernal *Support Vector Machine* (SVM) (Praghakusma & Charibaldi, 2021). Penelitian ini diawali dengan pengambilan data dari media sosial Twitter dan Instagram yang menandai akun resmi KPK, dilanjutkan dengan proses labeling data secara manual, *preprocessing* data, pembobotan, hingga pengujian. Untuk mengetahui opini masyarakat terhadap institusi KPK, penelitian ini secara eksplisit berupaya menciptakan sistem cerdas

yang dapat mengkategorikan komentar-komentar yang disampaikan masyarakat di Instagram dan Twitter. Berdasarkan percobaan komparasi yang dilakukan, diketahui bahwa kernel yang memiliki nilai akurasi paling tinggi adalah kernel linier, yakni sebesar 83,06% diikuti oleh kernel polynomial dengan akurasi sebesar 81,45%, dan kernel sigmoid yang menghasilkan nilai akurasi sebesar 79,83%.

Dalam salah satu penelitian, Oryza Habibie Rahman dan kawan-kawan melakukan klasifikasi terhadap ujaran kebencian yang ada pada media sosial twitter menggunakan SVM (Rahman *et al.*, 2021). Proses penelitian melibatkan pengumpulan data dari Twitter, praproses data dengan tahapan seperti case folding, filtering, tokenizing, stemming, dan pembobotan menggunakan TF-IDF, serta klasifikasi menggunakan SVM. Dalam penelitian ini, tiga buah kernel digunakan untuk klasifikasi ujaran kebencian, yaitu kernel linear, sigmoid, dan RBF. Hasil pengujian menunjukkan bahwa kernel RBF memiliki akurasi tertinggi sebesar 93%, presisi 84%, recall 86%, dan F-measure 83% . Kernel linear memiliki akurasi sebesar 92%, presisi 85%, recall 88%, dan F-measure 85% . Sedangkan kernel sigmoid memiliki akurasi sebesar 92%, presisi 90%, recall 89%, dan F-measure 87% . Dengan demikian, kernel RBF menunjukkan performa yang lebih baik dalam klasifikasi ujaran kebencian dibandingkan dengan kernel linear dan sigmoid.

Penelitian mengenai metode *Support Vector Machine* (SVM) untuk analisis sentimen juga dilakukan oleh Setyawati dan kawan-kawan tahun 2021. Penelitian yang dilakukan adalah menganalisis pendapat masyarakat terhadap Program Kartu Prakerja menggunakan metode Support Vector Machine (SVM) (Hendrastuty *et al.*, 2021). Data yang digunakan pada penelitian ini adalah data sekunder berupa teks

berbahasa Indonesia yang diperoleh dari media sosial X atau Twitter. Untuk mengukur kinerja dari klasifikasi SVM yang telah dibuat, digunakan *Confusion Matrix.* Pada penelitian ini juga dilakukan perbandingan antara kernel linear dan kernel RBF. Hasil penelitian menunjukkan bahwa kernel linear lebih unggul daripada kernel RBF berdasarkan tingkat akurasi yang dihasilkan. Kesimpulan dari penelitian ini adalah bahwa sentimen masyarakat cenderung netral terhadap program kartu prakerja. Berdasarkan hasil evaluasi, akurasi kernel linear sebesar 98.67%, sedangkan akurasi kernel RBF sebesar 98.34%. Dari hasil tersebut, dapat disimpulkan bahwa kernel linear memiliki tingkat akurasi yang sedikit lebih tinggi daripada kernel RBF dalam mengklasifikasikan sentimen masyarakat terhadap Program Kartu Prakerja.

Pada tahun 2023, Rani Yunita dan Mia Kamayani meneliti perbanidngan antara penggunaan Agoritma SVM dengan Naïve Bayes untuk analisis sentimen terhadap kebijakan penghapusan kewajiban skripsi (Yunita & Kamayani, 2023). Penelitian yang dilakukan oleh dua peneliti pada penelitian ini bertujuan untuk melakukan analisis sentimen terhadap kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan di perguruan tinggi di Indonesia . Proses penelitiannya melibatkan pengumpulan dataset dari Twitter menggunakan metode tweet-harvest, filtering dataset, preprocessing dataset dengan Rapidminer, serta melakukan analisis sentimen dengan menggunakan algoritma SVM dan Naïve Bayes. Berdasarkan penelitan yang dilakukan, didapatkan nilai akurasi untuk masing-masing algoritma SVM dan Niave Bayes adalah 80% dan 75%. Lebih jelasnya, hasil penelitian menunjukkan bahwa SVM merupakan algoritma terbaik dengan

akurasi 80%, recall 83%, presisi 76%, dan F1-Score 79%, sedangkan Naïve Bayes, diperoleh nilai akurasi sebesar 75%, recall sebesar 75%, presisi sebesar 72%, dan F1-Score sebesar 73%.

Tabel 2. 1 Penelitian Terkait

No	Referensi	Input	Metode	Hasil
1.	Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter (Alhaq *et al*., 2021)	- Dataset yang digunakan merupakan tweet terkait Bukalapak yang dikumpulkan menggunakan Twitter Scrapper dan disimpan dengan bentuk .csv. - Tiga kategori label—positif, negatif, dan netral—diterapkan pada kumpulan data yang diberi label secara manual	- *Support Vector Machine* - Preprocessing data dilakukan meliputi tahap cleaning, tokenizing, normalisasi, stopword removal, POS tagging, POS filtering, dan stemming.	Hasil klasifikasi sentimen yang dilakukan dengan SVM mencapai akurasi tertinggi sebesar 93% dengan metode K-Fold Cross Validation
2.	Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi (Praghakusma & Charibaldi, 2021)	- Dataset yang digunakan diperoleh melalui metode *scrapping* dari *tweet* yang menandai akun resmi KPK dan komentar pada kolom komentar akun KPK di instagram. - Dataset yang berjumlah 1614 terbagi menjadi tiga label kategori, yaitu positif dengan jumlah 538 data, negatif 538 data, dan netral berjumlah 538 data.	- *Support Vector Machine* - Preprocessing data dilakukan meliputi tahap case folding, cleansing, tokenizing, stopword removal, dan stemming.	Dari komparasi kernel yang dilakukan, diketahui bahwa kernel yang memiliki nilai akurasi paling tinggi adalah kernel linier, yakni sebesar 83,06% diikuti oleh kernel polynomial dengan akurasi sebesar 81,45%, dan kernel sigmoid yang menghasilkan nilai akurasi sebesar 79,83%.
3.	Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine (Rahman *et al*., 2021).	- Data yang digunakan merupakan data *tweet berupa kalimat* ujaran kebencian yang terdiri dari kelas suku, agama, ras, antar golongan, dan netral bermformat .csv. Data diperoleh melalui proses *crawling.*	- *Support Vector Machine* - Preprocessing yang dilakukan terdiri dari *case folding, tikenizing, filtering,* dan *stemming.*	Dalam penelitian ini menggunakan 3 kernel untuk pengujiannya dan . hasil pengujian menunjukkan bahwa kernel RBF memiliki akurasi tertinggi sebesar 93%, presisi 84%, recall 86%, dan F-measure 83% . Kernel linear memiliki

No	Referensi	Input	Metode	Hasil
				akurasi sebesar 92%, presisi 85%, recall 88%, dan F-measure 85% . Sedangkan kernel sigmoid memiliki akurasi sebesar 92%, presisi 90%, recall 89%, dan F-measure 87% .
4.	Analisis Sentimen Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine (Styawati *et al*., 2021)	- Data dipeoleh melalui sosial media *Tweeter* dengan kata kunci "Prakerja" dalam rentang waktu 22 April – 29 April 2021 - Data yang digunakan berjumlah 2000 data	- *Support Vector Machine* - Proses preprocessing yang dilakukan pada data adalah *cleansing, case folding, tokenizing, filtering,* dan *stemming.* - Pembobotan pada data menggunakan TF-IDF	Dilakukan perbandingan antara dua kernel, yaitu kernel lineardan RBF. Berdasarkan hasil evaluasi, akurasi kernel linear sebesar 98.67%, sedangkan akurasi kernel RBF sebesar 98.34%.
5.	Analisis Sentimen Terhadap Kebijakan Penghapusan Kewajiban Skripsi (Yunita & Kamayani, 2023)	- Data yang digunakan merupakan *tweet* dengan kata kunci "Nadiem skripsi", "lulus tanpa skripsi", "pengganti sktipsi", dan "skripsi dihapus" berjumlah 1303 data komentar - Kategori sentimen negatif dan positif dibuat secara manual dari data.. Perbandingan masing-masing kategori adalah 331 untuk sentiment positif dan 369 untuk sentiment negatif.	- *Support Vector Machine* dan Naïve Bayes - Proses filtering dilakukan untuk memilih data yang akan digunakan pada tahap selanjutnya, sehingga diperoleh 700 data dari data awal. - Preprocessing yang dilakukan pada dataset adalah *cleansing, transform cases, tokenizing, filter stopword,* dan *filter token (by length).*	Hasil penelitian menunjukkan bahwa SVM merupakan algoritma terbaik dengan akurasi 80%, recall 83%, presisi 76%, dan F1-Score 79%, sedangkan Naïve Bayes, diperoleh nilai akurasi sebesar 75%, recall sebesar 75%, presisi sebesar 72%, dan F1-Score sebesar 73%.

Penelitian terdahulu yang disajikan pada tabel 2.1 menjadi bukti bahwa metode *Support Vector Machine* (SVM) mampu melakukan analisis sentimen

dengan baik. Oleh karena itu, metode SVM dipilih untuk melakukan analisis sentimen pada data yang belum pernah digunakan dalam data penelitian terdahulu, yaitu ulasan produk Cetaphil Gentle Skin Cleanser yang diperoleh melalui website Female Daily. Dengan mengacu pada penelitian terdahulu, penelitian ini akan menguji kombinasi proses *preprocessing* yang digunakan dalam analisis sentimen. Dalam kondisi ini, pengujian yang dilakukan adalah pengaruh proses *normalization* pada *preprocessing.* Namun, sebelum melakukan hal tersebut, pengujian terhadap rasion data *train* dan *testing* akan dilakukan untuk mengetahui rasio yang memiliki kinerja paling baik dalam penelitian ini. Selain itu, mengacu pada penelitian nomor 3 dan 4, perbandingan kinerja kernel dalam SVM akan dilakukan untuk mengtahui kernel yang memiliki kinerja terbaik.

2.2 Analisis Sentimen

Analisis sentimen atau juga dikenal dengan *opinion mining* didefinisikan sebagai proses memahami serta mengolah data berupa teks untuk mendapat informasi sentimen yang terkandung dalam kalimat opini yang ada. Ini berguna untuk menemukan informasi penting dari data yang tidak terstruktur (Giovani *et al*., 2020). Analisis sentimen digunakan untuk menentukan polaritas atau kekuatan opini (positif atau negatif) yang diungkapkan dalam teks tertulis (Taj *et al*., 2019). Tujuan analisis ini adalah untuk mengklasifikasikan sentimen kemudian memberikan label kepada teks dalam beberapa kategori (Araque *et al*., 2019).

Tugas utama dari analisis sentimen ialah untuk memahami sekaligus mengelompokkan emosi dari suatu teks yang berbentuk opini, kalimat, ataupun dokumen. Oleh karena itu, mengetahui pemikiran masyarakat terhadap suatu

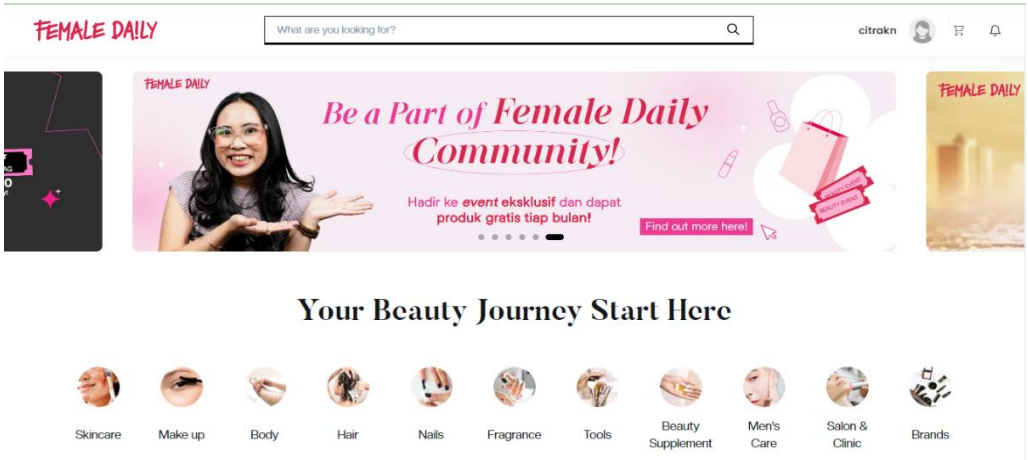
barang atau jasa melalui analisis sentimen akan memberikan manfaat bagi banyak pihak. Karena dapat dijadikan bahan penilaian maka analisis sentimen menjadi sangat penting.

Data sentimen yang akan dianalisis harus bebas dari berbagai jenis gangguan atau *noise* sebelum analisis sentimen dapat dilakukan Untuk itu akan dilakukan *preprocessing* guna memperoleh data yang bersih sebelum dilakukan analisis sentimen. Langkah yang digunakan untuk membersihkan dan mengatur teks yang belum diproses disebut *preprocessing*. Secara umum, pembersihan teks ini dilakukan dalam beberapa tahap, yaitu *cleansing* untuk menghapus karakter-karakter khusus dalam teks, pengubahan teks menjadi huruf kecil semua (*case folding*), penghapusan kata-kata yang tidak memberikan informasi penting, dan melakukan *stemming* untuk mengubah kata-kata menjadi bentuk dasarnya.

Penelitian terkait pegaruh *preprocessing* pernah dilakukan oleh Syifa Khairunnisa dan kawan-kawan pada tahun 2021. Penelitian ini meneliti tentang pengaruh *text preprocessing* terhadap analisis sentimen dan mengambil studi kasus pada komentar masyarakat terkait pandemic Covid-19 (Khairunnisa *et al*., 2021). Melalui penelitian ini, didapatkan bahwa kombinasi *preprocessing* memiliki pengaruh besar dalam melakukan analisis sentimen untuk menghasilkan nilai akurasi yang besar. Dari beberapa scenario yang dilakukan, didapatkan hasil bahwa kombinasi yang menghasilkan nilai akurasi terbesar yaitu propses *preprocessing* yang hanya menlibatkan *cleaning* dan *stemming* (tanpa adanya normalisasi kata maupun *stopword removal*) serta kombinasi *preprocessing* yang terdiri dari normalisasi, *cleaning*, dan *stemming* (tanpa disertai *stopword removal*).

2.3 *Female Daily*

Di era digital ini, pengguna media social tidak lagi terbatas pada individu saja, melainkan suatu perusahaan atau bisnis juga. Tak jarang suatu bisnis atau perusahaan memanfaatkan platform media untuk memelibatkan pelanggan dalam diskusi produksi, inovasi, pemasaran, hingga berbagi informasi (Madiistriyatno & Alwiyah, 2023), contohnya adalah Female Daily. Melalui website resmi femaledaily.com, dijelaskan bahwa Female Daily bermula dari forum sederhana yang anggotanya dianggap sebagai pengguna baru produk kecantikan kemudian berkembang menjadi salah satu komunitas terbesar yang berpusat pada perempuan Indonesia. Female Daily merupakan platform yang menyediakan wadah bagi penggunanya untuk berbagi pengalaman ketika menggunakan produk kecantikan (Francia & Salman. 2022). Platform ini ramai digunakan untuk melihat ulasan produk serta melakukan diskusi secara intens antar penggunanya (Atthahahirah, 2023). Tampilan website Female Daily disajikan pada gambar 2.1

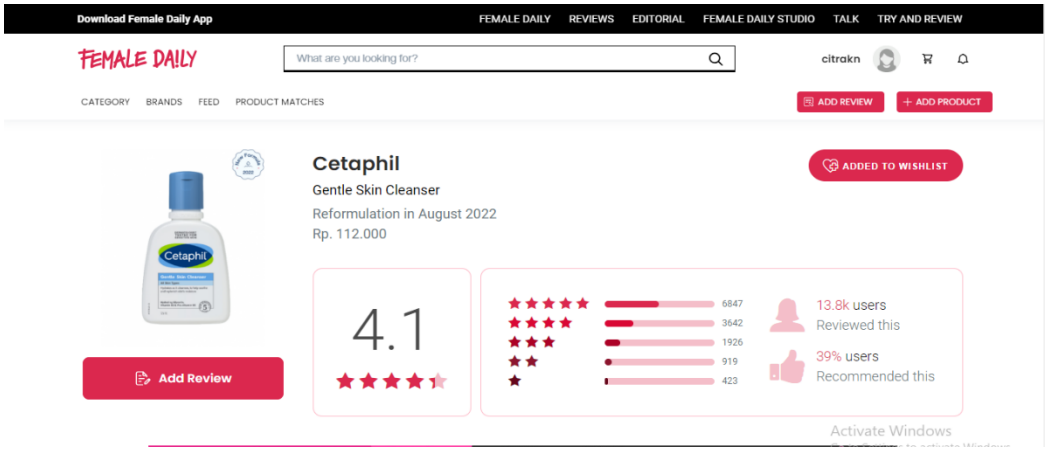


Gambar 2.1 Tampilan Website Female Daily

Gambar yang disajikan pada gambar 2.1 adalah tampilan utama dari website Female Daily yang dapat diakses melalui “femaledaily.com.” Saat ini Female Daily tidak hanya menjadi tempat untuk berbagi pengalaman dalam penggunaan *skincare* melalui ulasan yang diunggah, tapi berbagai produk kecantikan lainnya seperti *make up,* parfum, hingga berbagai produk untuk laki-laki.

2.4 Cetaphil Gentle Skin Cleanser

Cetaphil merupakan salah satu *brand* yang tersebar di Indonesia. Dibuat oleh seorang apoteker pada tahun 1947 di Texas, eksistensi Cetaphil sebagai *brand* yang dikenal sebagai *brand* ilmu kulit sensitive tidak pudar sedikitpun. Cetaphil telah hadir di Indonesia sejak tahun 2013 dan banyak direkomendasikan oleh ahli dermatologi karena memiliki formula yang lembut dan sangat minim iritasi. Salah satu produk dari *brand* ini adalah Gentle Skin Cleanser. Produk yang dirilis dengan nama Cleansing Lotion ini merupakan produk pertama Cetaphil yang masih eksis sampai sekarang, bahkan telah terjual lebih dari 10 ribu buah di salah satu e-commerce. Dilansir dari website resmi Cetaphil, formula gel yang dibuat untuk produk ini terbukti secara klinis memberikan hidrasi yang terus-menerus untuk melindungi agar tetap lembab. Berikut ini ditampilkan laman review produk Cetaphil Gentle Cleanser di Female Daily.Gambar 2.2 merupakan tampilan salah satu produk Cetaphil yang diunggah untuk diberi ulasan. Produk Cetaphil Gentle Skin Cleanser telah diulas oleh belasan ribu orang dengan rating 4,1.



Gambar 2.2 Tampilan Halaman Review Produk Cetaphil

2.5 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Proses *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah salah satu metode pembobotan yang umum digunakan dalam perolehan informasi dan text mining (Artama *et al*., 2020). Proses TF-IDF bertujuan untuk menghasilkan representasi vektor yang menggambarkan tingkat kepentingan setiap kata dalam dokumen terhadap keseluruhan korpus. Proses perhitungan TF-IDF membantu memfilter dan menonjolkan kata-kata kunci yang memiliki kontribusi tinggi terhadap makna dan sentimen dalam setiap dokumen.

Semakin sering sebuah kata atau frase muncul dalam sebuah teks, maka semakin tinggi pula skor yang diperoleh algoritma TF-IDF. Setiap kata yang ada akan memiliki nilai yang ditetapkan padanya. Nilai-nilai ini kemudian akan diurutkan dari yang terbesar ke terkecil atau sebaliknya. Hasilnya adalah representasi numerik yang memungkinkan analisis sentimen yang lebih akurat dengan mempertimbangkan bobot relatif kata-kata dalam konteks yang telah ditentukan. Pembobotan ditentukan dengan dengan mengalikan *term frequency*

(TF)—frekuensi kemunculan suatu kata dalam sebuah dokumen—dengan inverse document frekuensi (IDF), yaitu jumlah kemunculan kata dalam kumpulan dokumen.

Berikut ini adalah rumus yang digunakan dalam menghitung nilai *Inverse Document Frekuensi* (IDF):

$$IDF \quad = \log\left(\frac{N}{ni}\right) \tag{2.1}$$

Untuk perhitungan TFIDF dapat dilakukan menggunakan persamaan (2.2) berikut

$$TFIDF = TF \times IDF \tag{2.2}$$

Dimana:
TF = Jumlah kemunculan (i) term pada sebuah dokumen
IDF = Jumlah (i) term pada seluruh dokumen
N = Jumlah dokumen
ni = Jumlah dokumen yang mengandung (i) term

2.6 *Support Vector Machine* (SVM)

Untuk klasifikasi teks otomatis, salah satu jenis teknik pembelajaran supervised yang dapat digunakan adalah algoritma Support Vector Machine. (Luqyana, 2018). Algoritma ini secara sederhana dapat dipahami sebagai upaya untuk memaksimalkan jarak antara dua kelas guna mengidentifikasi hyperplane optimal yang berfungsi sebagai pemisah antara keduanya dalam ruang input. (Alhaq *et al*., 2021). Keunggulan dari algoritma ini terletak pada kemampuannya untuk mengimplementasikan pemisahan linear pada input data yang memiliki dimensi yang lebih besar.

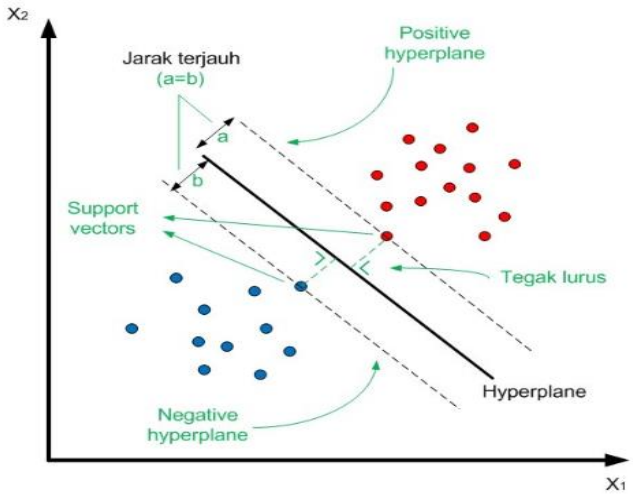
Muhammad Samantri dan Afiya pada penelitian terbaru tahun 2024 meneliti perbandingan antara algoritma *Support Vector Machine* dan *Random Forest* untuk analisis sentimen dengan mengambil studi kasus terhadap kebijakan pemerintah terkait kenaikan harga BBM yang terjadi pada tahun 2022 (Samantri & Afiyati., 2024). Data uang digunakan dalam penelitian ini adalah data twitter yang dikumpulkan menggunakan API dan library Tweepy. Hasil dari penelitian yang dilakukan oleh Muhammad Samantri dan Afiyanti menunjukkan bahwa SVM memiliki tingkat akurasi lebih tinggi yakni 77% dibandingkan dengan *Random Forest* yang menghasilkan tingkat akurasi 76%.

Penlitian terkait perbandingan algoritma SVM dan yang lainnya dilakukan oleh Chindy Aulia Sari da kawan-kawan pada tahun 2022. Pada penelitian ini dilakukan perbandingan antara 3 metode, yaitu Naïve Bayes, SVM, dan *Decision Tree* untuk klasifikasi konsumsi obat. Pada penelitian ini, pembagian data uji dan data latih dilakukan secara random dengan jumlah 20% dari keseluruhan baris dataset. Dari perbandingan tersebut, didapatkan hasil yang menunjukkan bahwa metode SVM memiliki nilai akurasi paling tinggi sebesar 53%, diikuti dengan Naïve Bayes 50%, dan *Decision Tree* 41%.

Penelitian yang dilakukan penulis menerapkan metode SVM untuk melakukan analisis sentimen. Secara konseptual, Untuk membagi dua kelas data secara efektif, SVM berupaya mengidentifikasi hyperplane yang optimal. Untuk membagi dua kelas data secara efektif, SVM berupaya mengidentifikasi hyperplane yang optimal. Dua kelas dibagi oleh hyperplane, yang merupakan bidang data penentu n-dimensi..Nilai *hyperlane* terbaik didapat dengan cara mengukur margin

yang merupakan jarak terjauh antara *hyperlane* dengan *pattern* paling dekat (*support* vector) dari setiap kelas.

Contohnya data pelatihan ditulis sebagai (x_i, y_i) di mana $i = 1, 2, 3, \dots, n. x_i = [x_{i1}, x_{i2}, \dots x_{ij}]$ adalah vektor baris dari fitur ke-i di ruang dimensi ke-j dan yi merupakan label dari x_i yang diasumsikan sebagai $y_i \in \{-1, +1\}$. Dua kelas -1 dan +1 diasumsikan dipisahkan secara linear oleh sebuah *hyperplane.* Pada gambar 2.3, *hyperlane* ditunjukkan dengam garis hitam yang ada ditengah. Kelas +1 pada gambar merupakan kelas positif dan ditunjukkan oleh data yang terletak di atas *hyperlane* dan sebaliknya data yang terletak di bawah *hyperlane* adalah kelas -1.



Gambar 2.3 Ilustrasi Support Vector Machine

Untuk mencari nilai *hyperlane,* dapat menggunakan persamaan berikut ini:

$$f(x) = w.xi + b \quad (2.3)$$

Dimana:
w = Parameter bobot
xi = Vector input
b = bias

Hyperplane tegak lurus terhadap vektor w.. Dengan kata lain, hyperplane juga akan mengikuti jika nilai b berubah.. Oleh karena itu, diperlukan pencarian nilai margin terbesar untuk mengidentifikasi hyperplane yang optimal. Sementara itu, jarak antara hyperplane dengan titik terdekatnya dimaksimalkan untuk menghasilkan nilai margin tertinggi.. *Pattern* atau pola yang memenuhi persamaan $w.xi + b \leq -1$ merupakan pola yang dapat memenuhi kelas -1 dan termasuk dalam kelas negatif.

Hyperplane pada SVM yang memisahkan dua kelas dalam ruang vitur biasanya digambarkan sebagai titik (x,y) *hyperlane* sebagai berikut:

$$Ax + By + C = 0 \qquad (2.4)$$

Dengan rumus jarak didapatkan melalui persamaan berikut ini:

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}$$

Persamaan 2.4 selanjutnya akan diubah menjadi *dot product* pada vektor sehingga menjadi seperti berikut:

$$[A\ B] \begin{bmatrix} x \\ y \end{bmatrix} + C = 0$$

Apabila $w = [A\ B]$, $xi = \begin{bmatrix} x \\ y \end{bmatrix}$ dan $b = C$, maka akan diperoleh persamaan berikut:

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}} = \frac{|w.xi + b|}{\sqrt{w^2 + c^2}} = \frac{|w.xi + b|}{||w||}$$

Nilai margin ditentukan dengan memanfaatkan jarak antar kelas, lebih tepatnya adalah nilai tengah dengan menerapkan rumus:

$$margin = \frac{1}{2}(d^{+} - d^{-}) \quad (2.5)$$

$$= \frac{1}{2}\left(\frac{|w.x_1 + b|}{||w||} - \frac{|w.x_2 + b|}{||w||}\right)$$

$$= \frac{1}{2}\left(\frac{1}{||w||} - \frac{-1}{||w||}\right)$$

$$= \frac{1}{||w||}, ||w|| \neq 0$$

Dengan:
d^{+} = Jarak *hyperlane* terhadap kelas +1 atau positif
d^{-} = Jarak *hyperlane* terhadap kelas -1 atau negatif

Nilai $||w||^2$ harus diminimalkan untuk memaksimalkan nilai margin yang sebanding. Dengan demikian, perhitungan hyperlane optimal dengan nilai limit terbesar dapat dirumuskan sebagai berikut:

$$\max margin = \frac{1}{2}||w||^2 \quad (2.6)$$

Persamaan tersebut dapat diubah menjadi *Lagrange* untuk menyelesaikan masalah ini.

$$min\ L_p(w, b, ai) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i\,[y_i\,(w.xi + b) - 1]$$

Dengan:
L_p := Fungsi *lagrange* (primal problem)
a_i = Koefisien *lagrange*, $a_i \geq 0$ dengan $i = 1, 2, \ldots, n$

Setelah meminimalkan L_p menjadi ai dan mereduksinya menjadi nilai b dan w, akan ditemukan turunan kreduksi terhadap nilai w dan b dan dimaksimalkan terhadap nilai ai lalu turunan awal fungsi L_p terhadap nilai b dan w, yang menghasilkan hasil seperti persamaan berikut:

Turunan awal dari fungsi L_p terhadap variabel w

$$\frac{\partial}{\partial w} L_p(w, b, a) = 0$$

Maka didapatkan persamaan nilai w seperti berikut:

$$min\ L_p(w, b, a) = \frac{1}{2}\ ||w||^2 - \sum_{i=1}^{n} a_i\ [y_i\ (w.xi + b)] + \sum_{i=1}^{n} a_i$$

$$\frac{\partial}{\partial w} L_p(w, b, a) = w - \sum_{i=1}^{n} a_i y_i . x_i$$

$$w = \sum_{i=1}^{n} a_i y_i . x_i \qquad (2.7)$$

Turunan pertama fungsi L_p terhadap nilai b

$$\frac{\partial}{\partial w} L_p(w, b, a) = 0$$

Maka didapatkan:

$$0 = \sum_{i=1}^{n} a_i y_i . x_i$$

Selanjutnya formula L_p diubah menjadi L_D (dual problem).

$$maxL_D(a) = \frac{1}{2} \left(\sum_{i=1}^{n} a_i y_i . x_i \right) \left(\sum_{i=1}^{n} a_i y_i . x_i \right)$$

$$- \sum_{i=1}^{n} a_i y_i \left(\left(\sum_{i=1}^{n} a_i y_i . x_i \right) x_i + b \right) + a_i$$

$$= \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i y_i a_j y_j (x_i\, x_j), \qquad (2.8)$$

Dengan kendala,

$$\sum_{i=1}^{n} a_i y_i = 0, a_i > 0$$

Persamaan (2.9) digunakan untuk menentukan nilai a_i menggunakan teknik substitusi. Nilai a_i ini berguna untuk mencari nilai w. Nilai a_i yang bernilai 0 tidak terlibat sama sekali dalam proses peramalan data baru. Hal ini karena nilai tersebut tidak akan dimasukkan dalam perhitungan untuk menentukan nilai w. Sedangkan untuk $a_i > 0$ memiliki sebutan *support vector*.

Selanjutnya persamaan (2.8) disubstitusikan ke persamaan (2.7) dengan menerapkan kernel linear $(x_i, x_j) = x_i . x_j$. untuk mengevaluasi data baru dengan model yang telah dilatih dengan melakukan $sign\{f(x)\}$, sehingga menghasilkan persamaan berikut:

$$f(x) = \sum_{i=1}^{n} a_i y_i (x_i^T . x) + b \tag{2.9}$$

Substitusikan persamaan di atas ke dalam $y_i f(x_i) = 1$, maka diperoleh:

$$y_i \sum_{m \in s} a_m \, y_m x_m^T + b = 1$$

Dengan
s = Himpunan Indeks Support Vector

Untuk menghitung b, digunakan persamaan:

$$y_i \left(\sum_{m \in s} a_m \, y_m x_m^T . x_i + b \right) = 1$$

$$y_i y_i \left(\sum_{i=m}^{s} a_m \, y_m x_m^T . x_i + b \right) = y_i$$

$$\left(\sum_{i=m}^{s} a_m \, y_m x_m^T . x_i + b \right) = y_i$$

$$b = y_i - \sum_{i=m}^{s} a_m \, y_m x_m^T . x_i$$

$$b = \frac{1}{N_s} \sum_{i \in s} \left(\sum_{i=m}^{s} a_m \, y_m x_m^T . x_i + b \right) \quad (2.10)$$

Dengan
N_s = Nilai Support Vector

Pendekatan umum untuk klasifikasi atau prediksi biasanya menggunakan *Support Vector Machine* (SVM). Hyperplane optimal untuk membagi dua kelas data adalah apa yang ingin diidentifikasi secara kontekstual oleh SVM. Pemisahan dua kelas ini dilakukan dengan meningkatkan margin atau jarak antar kelas data. Dengan memanfaatkan pendekatan kernel, algoritma ini dapat beroperasi pada dataset dengan dimensi tinggiKernel Radial Basic Function (RBF), kernel Linear, dan kernel Polynomial adalah beberapa jenis kernel yang sering digunakan di SVM Persamaan untuk seluruh kernel ditunjukkan pada tabel 2.2.

Tabel 2.2 Rumus Persamaan Fungsi Kernel

No	Kernel	Persamaan
1.	*Linear*	$K(x, y) = x.y$
2.	*Polynomial*	$K(x, y) = (x.y)^d$
3.	*RBF*	$K(x, y) = \exp(-y\|x - y\|^2)$

2.7 *Confusion Matrix*

Confusion Matrix adalah teknik yang digunakan untuk melihat performa pada model hasil analisis evaluasi. Teknik ini akan menghitung nilai *accuracy, precision, recall,* dan *f1 score* dari model *machine learning* yang telah dibangun. Terdapat empat kasus pada kondisi tertentu dalam perhitungan *confusion matrix.* Kondisi-kondisi tersebut disajikan pada tabel 2.3 (Melani *et al*., 2019).

Tabel 2.3 Kondisi Confusion Matrix

		Predict Values	
		1	0
Aktual Values	1	True Positif (TP)	False Negatif (FN)
	0	Flase Positif (FP)	True Negatif (TN)

True Positive (TP) merupakan kondisi saat hasil data aktuan dan hasil data prediksi benar. *True Negative* (TN) merupakan kondisi saat hasil data actual tidak benar dan hasil prediksi benar. *False Positive* (FP) merupakan kondisi ketika hasil prediksi salah sedangkan hasil data actual benar. *False Negative* (FN) merupakan kondisi dimana data hasil prediksi dan hasil data actual keduanya salah.

Accuracy adalah sebuah variabel yang diperoleh dengan membagi jumlah total kejadian dengan rasio temuan yang diklasifikasikan dengan benar. *Accuracy* merupakan persentase dari model yang telah dibuat. Semakin tinggi nilai persentase yang didapat, maka semakin baik pula model yang telah dibuat. Perhitungan *accuracy* dilakukan menggunakan persamaan berikut (Kurnianto & Febriawan, 2023).

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP} \times 100\% \quad (2.11)$$

Precision merupakan variable yang digunakan untuk menentukan kebenaran proposi prediksi kelas positif. *Precision* menunjukkan tingkat ketepatatan informasi yang digunakan dalam membentuk jawaban oleh system. Nilai *precision* dapat dihitung menggunakan persamaan berikut (Zulqornain & Adikara, 2021).

$$Precision = \frac{(TP)}{TP + FP} \times 100\% \quad (2.12)$$

Recall adalah variable yang berguna untuk menghitung nilai persentases dari total positif yang diprediksi positif dengan benar dan bertujuan untuk menghitung kebenaran proporsi prediksi actual yang salah diprediksi. *Recall* dapat dicari dengan menggunakan persamaan berikut (Zulqornain & Adikara, 2021).

$$Recall = \frac{(TP)}{TP + FN} \times 100\% \quad (2.13)$$

Perhitungan untuk *F1 score* melibakan perhitungan *precision* dan *recall* karena *f1 score* adalah sebuah nilai tunggal yang mewakili keduanya. Oleh karena itu, perhitungannya dapat dijadikan sebagai rata-rata harmonik dari *precision* dan *recall* dengan masing-masing diberi bobot yang sama. Nilai *F-Measure* atau *F1 score* dapat diperoleh menggunakan persamaan berikut (Zulqornain & Adikara, 2021).

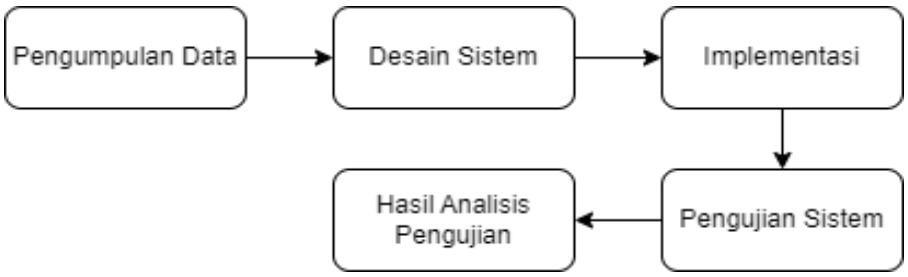
$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \quad (2.14)$$

BAB III

DESAIN DAN IMPLEMENTASI PENELITIAN

3.1 Prosedur Penelitian

Untuk memenuhi target penelitian, rancangan aktivitas yang sistematis tentu akan sangat berguna. Prosedur penelitian adalah istilah yang digunakan untuk menggambarkan rancangan aktivitas ini.. Bagian ini akan memberikan penjelasan setiap langkah proses penelitian, mulai dari pengumpulan data hingga hasil akhir. Prosedur yang akan digunakan dalam penelitian ini dijelaskan pada Gambar 3.1 di bawah ini:



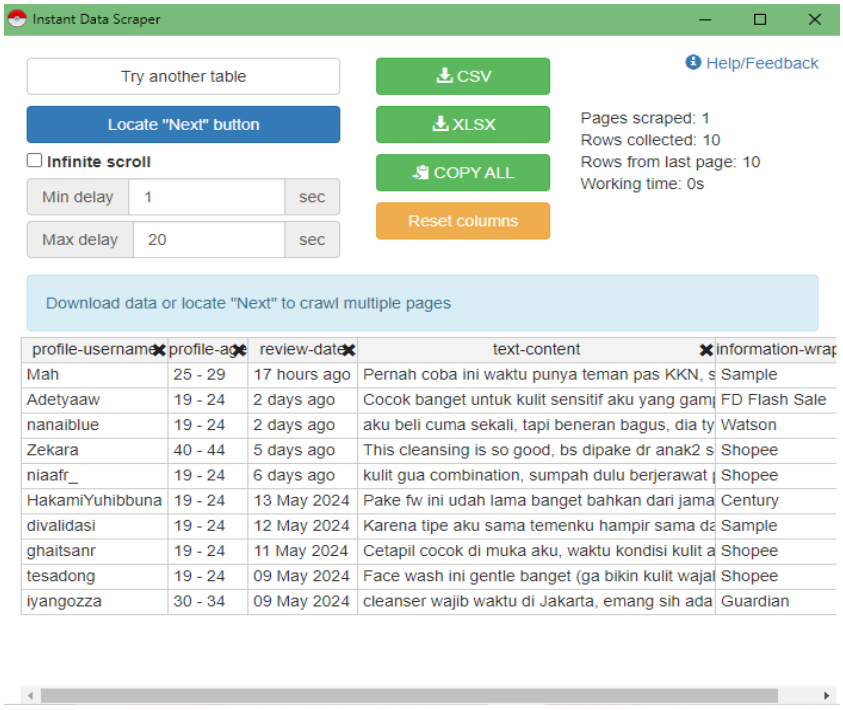
Gambar 3.1 Prosedur Penelitian

Rancangan aktivitas dalam prosedur penelitian ini dimulai dengan mengidentifikasi masalah sebagai dasar penelitian. Pada tahap ini juga dirumuskan rumusan masalah dan batasan masalah untuk menentukan fokus penelitian. Setelah itu, berdasarkan judul yang dipilih, referensi di berbagai tesis akademik dan jurnal ilmiah mulai dicari. Pengumpulan informasi penting lainnya yang diperlukan untuk penelitian adalah pengumpulan data. Dalam hal ini, data yang dikumpulkan didapatkan mengunakan tools *Instant Data Scrapper* yang tersedia di google. Tahap selanjutnya adalah desain sistem, di mana semua persyaratan terkait sistem akan

diuraikan secara rinci. Desain sistem yang dikembangkan berfungsi sebagai dasar pengujian dan implementasi. Setelah itu, sistem yang telah dibuat akan melewati tahap pengujian sistem, dan hasilnya nanti akan dianalisis untuk memastikan keefektifan dan keandalan implementasi.

3.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data primer berupa ulasan pengguna produk etaphil Gentle Skin Cleanser di platform Female Daily. Data ini diperoleh mengunakan tools *Instant Data Scrapper* yang tersedia di Google Chrome. Seluruh data yang diambil berjumlah 1000 data ulasan mulai dari tahun 2018 sampai dengan awal bulan Mei 2024. Hasil dari data ulasan selanjutnya disimpan dalam bentuk .csv. gambar 3.2 menampilkan proses pengambilan data ulasan menggunakan *Instant Data Scrapper.*



Gambar 3.2 Proses Pengambilan Data Ulasan Komentar

3.3 Pelabelan Data

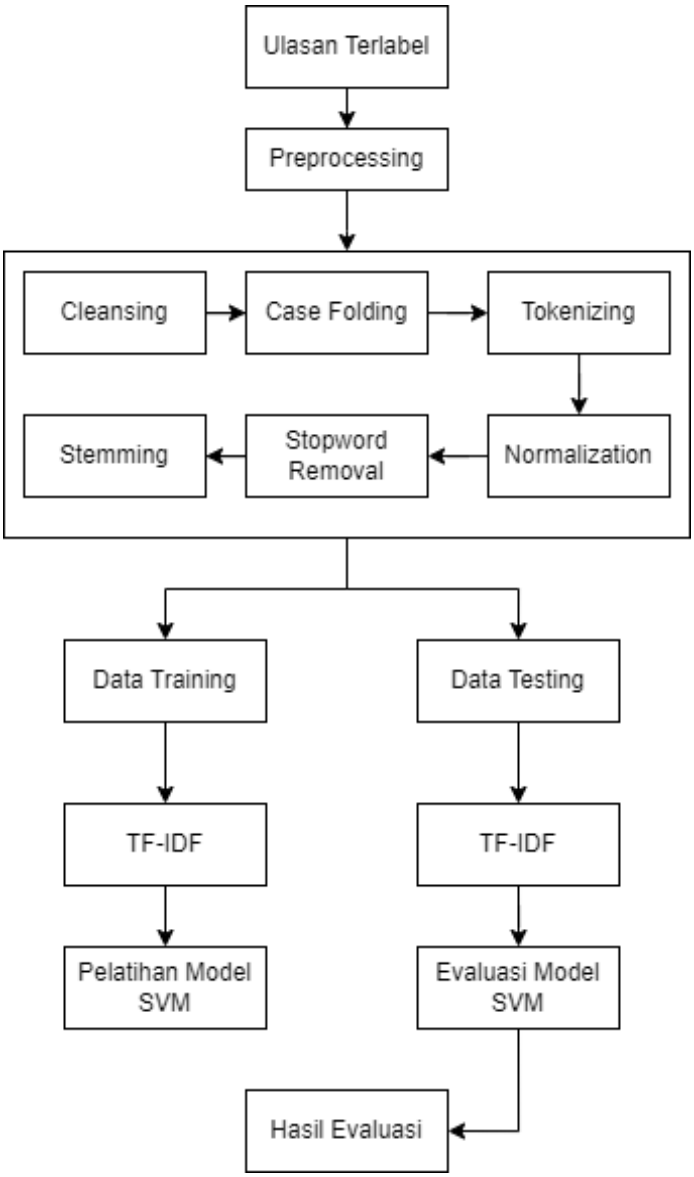
Data yang telah terkumpul selanjutnya akan diberi label. Tahap ini dilakukan untuk mengkategorikan data komentar menjadi dua kelompok: positif dan negatif. Ulasan yang dikategorikan positif adalah ulasan yang mengekspresikan kepuasan terhadap produk Cetaphil Gentle Skin Cleanser. Sebaliknya ulasan berlabel negatif adalah ulasan yang diberikan pengguna atas ketidakpuasan terhadap produk tersebut. Pelabelan data dilakukan secara manual oleh penulis kemudian untuk menghindari kesalahan selama pelabelan, hasil pelabelan akan diperiksa oleh validator. Validasi data dilakukan oleh Suciah Yastari, S.Pd yang mengajar di SMK Negeri 2 Selong, Lombok Timur sebagai guru Bahasa Indonesia. Hasil pelabelan data disajikan pada tabel 3.1.

Tabel 3.1 Ulasan Terlabel

No	Ulasan	Label
1	Cetaphil bagus banget bikin wajah bersih gaada bau sama sekali dan ga banyak busa jadi gak ganggu sama sekali tapi buat wajah bersih banget, rekomendasi banget buat wajah yang sukabruntusan dan kering, wajahku super sensitive cocok banget pake ini	positif
2	cetaphil? ini bagus banget aku udah cobaa dan bener kalau dia betul gentle skin cleanser yang ringan banget gaada rasa perih sehabis cuci muka gaada rasa cekat cekit dan mantep banget buat laki laki juga enak bagus banget dan buat kulit sensitif aman	positif
3	sabun cetaphil ini cocok banget untuk tipe kulit aku yg sensitive dan kering karena gentle dan tidak berbusa, jadi buat kalian yang punya kulit sensitive atau kering aku rekomendasiin sabun cetaphil inii yaaaa, smg membantu	positif
4	G tau kenapa dikulit aku malah g membantu menghilangkan jerawat mungkin karna menurut aku g ada busa atau aku pake g banyak jadi masih kurang bersih, aku juga g tauu. Yg jelas aku sedih ketika g bisa cocok sm sabun sejuta umat iniðŸ˜-	negatif
5	untuk tipe kulit wajah oily sepertiku kurang cocok sih langsung timbul jerawat, mungkin memang cocok pakai yang khusus oily, aku beli waktu yang oily kosong. Sedih sih yang lain kulit oily pake ini cocok, di aku sama sekali enggak, malah jerawat keluar terus, 1 bulan aku lanjutin dan akhirnya berhenti makin banyak jerawat yang timbul.	negatif
6	Aku ga cocok pake ini malah tumbuh jerawat batu dimana manaðŸ˜-terus kulanjutin eh makin parah mukaku,kujadiin sabun buat mandi eh badanku jadi ada jerawat sedih banget mana susah lagi ngilanginnya,emang ya skincare cocok cocokan yg cocok banget diorang tapi malah ga cocok diaku	negatif

3.4 Desain Sistem

Dalam penelitian ini, desain sistem dikembangkan sebagai alur sistem. Google Colab dan bahasa pemrograman Python digunakan untuk membangun sistem ini. Rancangan sistem secara lengkap disajikan pada Gambar 3.3 sebagai berikut:



Gambar 3.3 Desain Sistem

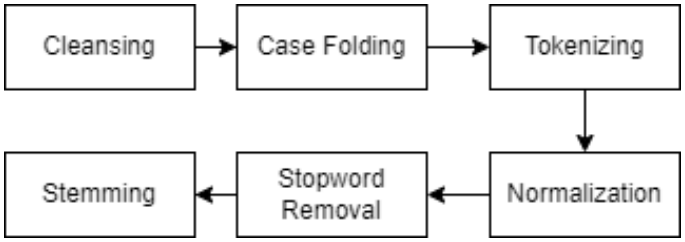
Alur perancangan sistem dalam penelitian ini diawali dengan tahap pengambilan data ulasan pengguna di female daily terkait produk Cetaphil Gentle Skin Cleanser, dilanjutkan dengan pemberian label atau sentiment pada setiap ulasan. Proses pegambilan data pada penelitian ini memanfaatkan sebuah *Extension* Google bernama *Instant Data Scrapper.*

Selanjutnya, data yang telah dikumpulkan akan melalui tahapan *preprocessing*. Tahap preprocessing dalam penelitian ini mencakup langkah-langkah seperti pembersihan (*cleansing*), pengubahan text menjadi huruf kecil semua (*case folding*), tokenisasi, *stopword removal*, dan *stemming.*

Hasil data dari tahap *preprocessing* akan dilanjutkan dengan pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) diterapkan guna mempersiapkan model dengan memanfaatkan representasi bobot kata untuk melatih dan menguji analisis sentimen. Pada tahap akhir penelitian, algoritma *Support Vector Machine* akan diterapkan untuk proses klasifikasi.

3.5 Text Preprocessing

Sebelum data diolah oleh model yang dikembangkan, terlebih dahulu harus melalui serangkaian prosedur yang disebut *preprocessing*. Langkah-langkah *preprocessing* penelitian ini ditampilkan dalam Gambar 3.4.



Gambar 3.4 Preprocessing

Tahap-tahap ini bertujuan untuk memastikan bahwa data yang digunakan dalam pelatihan maupun pengujian merupakan data yang representatif, bersih, serta siap untuk diproses. Selain itu, proses *preprocessing* ini akan mempengaruhi nilai performa proses klasifikasi sentimen. Enam langkah *preprocessing* yang digunakan dalam penelitian ini yaitu *cleansing*, *case folding*, tokenisasi, normalisasi, *stopword removal*, dan *stemming*.

3.5.1 *Cleansing*

Tahap *cleansing* melibatkan penghapusan elemen-elemen yang tidak diperlukan dalam teks. Tahap ini akan melibatkan penghapusan karakter dan tanda baca yang tidak relevan, termasuk koma, titik, tanda seru, dan tanda baca lainnya.. Selain itu, motikon dan karakter berlebihan lainnya dihilangkan, bersama dengan HTML dan URL, tagar, dan *mention*. Kode program untuk tahap ini ditampilkan seperti Gambar 3.5.

```
def cleanText(komentar):
    simbol = "!\"#$%&()*+-./:;<=>?@[\]^_`{|}~\n"
    komentar = re.sub("["
        u"\U0001F600-\U0001F64F"
        u"\U0001F300-\U0001F5FF"
        u"\U0001F680-\U0001F6FF"
        u"\U0001F1E0-\U0001F1FF"
    "]+", '', komentar, flags=re.UNICODE)
    komentar = re.sub(r'https*\S+', ' ', komentar)
    komentar = re.sub(r'@\S+', ' ', komentar)
    komentar = re.sub(r'rt', ' ', komentar)
    komentar = re.sub(r'#\S+', ' ', komentar)
    komentar = re.sub(r'\'\w+', '', komentar)
    komentar = re.sub(r'\w*\d+\w*', '', komentar)
    komentar = re.sub(r'\s{2,}', ' ', komentar)
    komentar = re.sub('rt', ' ', komentar)
    for i in simbol:
        komentar = komentar.replace(i, '')
    komentar = re.sub(r"#", " <hash_tag> ", komentar)
    komentar = re.sub(r'<[^<]+?>', '', komentar)
    komentar = komentar.replace('\n', ' ')
    komentar = re.sub(r'\s+', ' ', komentar)
    komentar = re.sub(r'\t', ' ', komentar)
    return komentar
```

Gambar 3.5 Kode Program Cleansing

Perbedaan antara ulasan sebelum dan sesudah dilakukan *cleansing* dapat dilihat pada tabel 3.2:

Tabel 3.2 Cleansing Ulasan

Sebelum Cleansing	Setelah Cleansing
Ini gak ada baunya sama sekali, dan aku seneng karna minim busa gak bikin kulit ketarik tapi kalo kalian suka type cleanser yang berbusa kau not rekomended sih dan ini cocok banget buat kulit yang lagi breakout yang butuh gak macem2 sama sekali	Ini gak ada baunya sama sekali dan aku seneng karna minim busa gak bikin kulit ketarik tapi kalo kalian suka type cleanser yang berbusa kau ga rekomended sih dan ini cocok banget buat kulit yang lagi breakout yang butuh gak sama sekali

3.5.2 *Case Folding*

Selanjutnya tahap ini akan dilakukan terhadap data yang telah melewati tahap *cleansing*. *Case folding* merupakan tahap di mana seluruh huruf yang terdapat dalam teks akan diubah menjadi huruf kecil. Sehingga hasil dari tahap ini berupa data bersih berisi kata-kata yang penting dan keseluruhan teksnya menjadi *lowercase*. Berikut ini merupakan kode program *Case Folding:*

```
ulasan['lower'] = ulasan['cleansing'].str.lower()
```

Gambar 3.6 Kode Program Case Folding

Pada tabel 3.3 dapat dilihat contoh perbedaan antara sebelum dilakukannya *cleaning* dan setelah dilakukannya *cleaning* pada sebuah ulasan :

Tabel 3.3 Case Folding Ulasan

Sebelum Case Folding	Setelah Case Folding
Ini gak ada baunya sama sekali dan aku seneng karna minim busa gak bikin kulit ketarik tapi kalo kalian suka type cleanser yang berbusa kau gak rekomended sih dan ini cocok banget buat kulit yang lagi breakout yang butuh gak sama sekali	ini gak ada baunya sama sekali dan aku seneng karna minim busa gak bikin kulit ketarik tapi kalo kalian suka type cleanser yang berbusa kau gak rekomended sih dan ini cocok banget buat kulit yang lagi breakout yang butuh gak sama sekali

3.5.3 *Tokenizing*

Tokenizing atau tokenisasi melibatkan pembagian teks menjadi token atau kata-kata individual. Tahap ini bertujuan untuk memecah teks menjadi bagian yang lebih kecil agar lebih mudah untuk dianalisis oleh model yang atau algoritma. Penerapan tokenizing sangat diperlukan dalam pembobotan kata TF-IDF karena prosesnya sendiri menggunakan token. Gambar 3.7 merupakan kode program *Tokenizing:*

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt')

def tokenizing(dokumen):
 dokumen = nltk.word_tokenize(dokumen)
 return(dokumen)
ulasan['tokenisasi'] = ulasan['lower'].fillna('').apply(lambda
x: tokenizing(x))
```

Gambar 3.7 Kode Program Tokenizing

Tabel 3.4 menyajikan perbandingan data yang telah menerapkan proses *tokenizing* dan belum menerapkannya pada ulasan.

Tabel 3. 4 Tokenizing Ulasan

Sebelum Tokenizing	Setelah Tokenizing
Ini gak ada baunya sama sekali dan aku seneng karna minim busa gak bikin kulit ketarik tapi kalo kalian suka type cleanser yang berbusa kau gak rekomended sih dan ini cocok banget buat kulit yang lagi breakout yang butuh gak sama sekali	['ini', 'gak', 'ada', 'baunya', 'sama', 'sekali', 'dan', 'aku', 'seneng', 'karna', 'minim', 'busa', 'gak', 'bikin', 'kulit', 'ketarik', 'tapi', 'kalo', 'kalian', 'suka', 'type', 'cleanser', 'yang', 'berbusa', 'kau', 'gak', 'rekomended', 'sih', 'dan', 'ini', 'cocok', 'banget', 'buat', 'kulit', 'yang', 'lagi', 'breakout', 'yang', 'butuh', 'gak', 'sama', 'sekali']

3.5.4 Normalization

Tahap *normalization* merupakan tahap untuk mengganti seluruh kata dalam teks ke dalam bahasa baku. Normalisasi memiliki tujuan untuk membuat data

lebih konsisten dan mengurangi variasi, sehingga mempermudah proses dalam analisis sentimen. Gambar 3.8 merupakan kode program *Normalization:*

```
normalized_word = pd.read_csv("drive/MyDrive/SKRIPSI
ICHA/normalisasi.csv", encoding = "latin1")

normalized_word_dict = {}
for index, row in normalized_word.iterrows():
    if row[0] not in normalized_word_dict:
        normalized_word_dict[row[0]] = row[1]

def normalized_term(document):
    return [normalized_word_dict[term] if term in
normalized_word_dict else term for term in document]

ulasan["normalisasi"] =
ulasan["tokenisasi"].apply(normalized_term)
```

Gambar 3.8 Kode Program Normalization

Tabel 3.5 menyajikan perbandingan data yang telah menerapkan proses normalisasi dan belum menerapkannya pada ulasan.

Tabel 3.5 Normalisasi Ulasan

Sebelum Normalisasi	Setelah Normalisasi
['ini', **'gak'**, 'ada', 'baunya', 'sama', 'sekali', 'dan', **'aku'**, 'seneng', **'karna'**, 'minim', 'busa', **'gak'**, 'bikin', 'kulit', 'ketarik', 'tapi', 'kalo', 'kalian', 'suka', 'type', 'cleanser', 'yang', 'berbusa', 'kau', 'gak', 'rekomendasi', 'sih', 'dan', 'ini', 'cocok', 'banget', 'buat', 'kulit', 'yang', 'lagi', 'breakout', 'yang', 'butuh', 'gak', 'sama', 'sekali']	['ini', **'tidak',** 'ada', 'baunya', 'sama', 'sekali', 'dan', **'saya'**, **'senang',** **'karena'**, 'minim', 'busa', **'tidak'**, **'buat'**, 'kulit', 'tarik', 'tetapi', 'kalau', 'kalian', 'suka', 'type', 'cleanser', 'yang', 'berbusa', 'kau', 'tidak', 'rekomendasi', 'sih', 'dan', 'ini', 'cocok', 'banget', 'buat', 'kulit', 'yang', 'lagi', 'breakout', 'yang', 'butuh', 'tidak', 'sama', 'sekali']

3.5.5 *Stopword Removal*

Stopword removal merupakan tahap yang melibatkan penghapusan kata-kata stop (stopwords) dalam teks. Kata-kata stop di sini maksudnya adalah kata-kata umum yang sering muncul dalam suatu bahasa, tetapi tidak memiliki kontribusi yang besar terhadap makna teks. Contoh kata-kata stop antara lain “dan”**,** “yang”,

“di”, “atau”, dan sebagainya. Berikut ini merupakan kode program *Stopword Removal:*

```
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')

def stopword_removal(dokumen):
 stop_word = set(stopwords.words('indonesian'))
 dokumen = [x for x in dokumen if x not in stop_word]
 return(dokumen)
ulasan['stopwords'] = ulasan['normalisasi'].apply(lambda x:
stopword_removal(x))
```

Gambar 3.9 Kode Program Stopword Removal

Tabel 3.6 menyajikan perbandingan data yang telah menerapkan proses *stopword removal* dan belum menerapkannya pada ulasan.

Tabel 3.6 Stopword Removal Ulasan

Sebelum Stopword Removal	Setelah Stopword Removal
['ini', **'tidak',** 'ada', 'baunya', 'sama', 'sekali', 'dan', **'saya'**, **'senang',** **'karena'**, 'minim', 'busa', **'tidak'**, **'buat'**, 'kulit', 'tarik', 'tetapi', 'kalau', 'kalian', 'suka', 'type', 'cleanser', 'yang', 'berbusa', 'kau', 'tidak', 'rekomendasi', 'sih', 'dan', 'ini', 'cocok', 'banget', 'buat', 'kulit', 'yang', 'lagi', 'breakout', 'yang', 'butuh', 'tidak', 'sama', 'sekali']	['tidak**,** 'baunya', 'senang', 'minim', 'busa', 'tidak**,** 'buat', 'kulit', 'tarik', 'suka', 'type', 'cleanser', 'berbusa', 'kau', 'tidak', 'rekomendasi', 'cocok', 'banget', 'kulit', 'breakout', 'butuh']

3.5.6 *Stemming*

Stemming merupakan proses penghaousan atau pemotongan imbuhan kata sehingga hanya menyisakan bentuk dasar dari kata tersebut. Tujuannya adalah untuk mengubah kata-kata yang berbeda, tetapi memiliki kata dasar yang sama menjadi bentuk yang seragam. Berikut ini merupakan kode program untuk *Stemming:*

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

def stemming(dokumen):
 factory = StemmerFactory()
 stemmer = factory.create_stemmer()
 dokumen = [stemmer.stem(word) for word in dokumen]
 dokumen = [t for t in dokumen if len(t) > 1]
 dokumen = " ".join(dokumen)
 return(dokumen)
ulasan['stemming'] = ulasan['stopwords'].apply(lambda x:
stemming(x))
```

Gambar 3.10 Kode Program Stemming

Tabel 3.7 menyajikan perbandingan data yang telah menerapkan proses *stemming* dan belum menerapkannya pada ulasan.

Tabel 3.7 Stemming Ulasan

Sebelum Stemming	Setelah Stemming
['tidak', **'baunya'**, 'senang', 'minim', 'busa', 'tidak', 'buat', 'kulit', 'ketarik', 'suka', 'type', 'cleanser', 'berbusa', 'kau', 'not', 'rekomended', 'cocok', 'banget', 'kulit', 'breakout', 'butuh']	['tidak', **'bau'**, 'senang', 'minim', 'busa', 'tidak', 'buat', 'kulit', 'tarik', 'suka', 'type', 'cleanser', 'busa', 'kau', 'tidak', 'rekomendasi', 'cocok', 'banget', 'kulit', 'breakout', 'butuh']

3.6 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Data yang telah melewati proses *preprocessing text,* selanjutnya akan memasuki proses *Term Frequency-Inverse Document Frequency* (TF-IDFPerhitungan bobot pada penelitian ini didasarkan pada seberapa sering suatu istilah muncul dalam suatu dokumen (TF) dan seberapa sering suatu istilah muncul di semua dokumen (IDF). Perhitungan bobot kata ini dilakukan menggunakan persamaan 2.1 dan 2.2 pada BAB II. Setelah nilai TF dan IDF diperoleh, maka dilanjutkan dengan pembobotan setiap *term* untuk memperoleh bobot masing-masing *term* dalam kumpulan dokumen. Tabel 3.8 berikut digunakan sebagai contoh perhitungan TFIDF.

Tabel 3.8 Data Sampel

Dokumen	Komentar
D1	cetaphil bantu ringan bruntusan wajah cocok kulit sensitif bruntusan kurang
D2	jerawat bruntusan pakai buat **jerawat** cepat kering tidak buat muka tarik gentle tidak bau busa sedikit tetap bersih wajah
D3	tidak cocok pakai produk buat kulit **jerawat**

Tahap pertama dalam menghitung bobot kata menggunakan TF-IDF adalah menghitung nilai TF itu sendiri. Tujuan perhitungan term pada setiap dokumen yang ada ialah untuk memperoleh frekuensi kata. Selanjutnya *df* dihitung untuk mengetahui banyaknya dokumen tempat suatu kata (term) tersebut muncul.

Nilai IDF dihitung menggunakan rumus pada BAB II setelah memperoleh nilai *df*. Contohnya kata "jerawat" berjumlah dua kata pada tabel 3.2, tepatnya pada D2 dan D3, sehingga nilai *df* = 2 dan banyaknya dokumen (N) = 3. Nilai IDF dihitung seperti berikut:

$$IDF = \log\left(\frac{N}{df}\right)$$

$$IDF = \log\left(\frac{3}{2}\right) = 0{,}176$$

Kemudian dengan persamaan 2.2, nilai TF-IDF dihitung dengan mengalikan jumlah term pada suatu dokumen (TF) dengan nilai IDF, sehingga didapatkan bahwa nilai TF-IDF pada D2 adalah sebagai berikut:

$$TFIDF = TF \times IDF$$

$$TFIDF = 2\ x\ 0{,}176 = 0{,}352$$

Sehingga kata "jerawat" pada D2 memiliki bobot yaitu 0,352.

Perhitungan pembobotan secara TF-IDF disajikan pada tabel 3.9 berikut:

Tabel 3.9 Hasil Pembobotan TF-IDF

	TF				IDF	TF-IDF		
Term	**D1**	**D2**	**D3**	**df**	**log (n/df)**	**D1**	**D2**	**D3**
cetaphil	1	0		1	0,477	0,477	0	0
bantu	1	0		1	0,477	0,477	0	0
ringan	1	0		1	0,477	0,477	0	0
beruntusan	1	1		2	0,176	0,176	0	0
wajah	1	1		2	0,176	0,176	0	0
cocok	1	0	1	2	0,176	0,176	0	0
kulit	1	0	1	2	0,176	0,176	0	0
sensitif	1	0		1	0,477	0,477	0	0,477
kurang	1	0		1	0,477	0,477	0	0,477
jerawat	0	2	1	2	0,176	0	0,352	0,176
pakai	0	1	1	2	0,176	0	0,176	0,176
buat	0	2	1	2	0,176	0	0,352	0,176
cepat	0	1		1	0,477	0	0,477	0
kering	0	1		1	0,477	0	0,477	0
tidak	0	2	1	2	0,176	0	0,352	0,176
muka	0	1		1	0,477	0	0,477	0
tarik	0	1		1	0,477	0	0,477	0
gentle	0	1		1	0,477	0	0,477	0
bau	0	1		1	0,477	0	0,477	0
busa	0	1		1	0,477	0	0,477	0
sedikit	0	1		1	0,477	0	0,477	0
tetap	0	1		1	0,477	0	0,477	0
bersih	0	1		1	0,477	0	0,477	0
produk	0	0	1	1	0,477	0	0	0,477

3.7 Implementasi Support Vector Machine (SVM)

Proses pelatihan pada implementasi metode SVM dilakukan dengan tujuan untuk menemukan nilai vector a, nilai w, dan bias (b). Ulasan yang positif akan ditandai dengan label 1 dalam prosedur ini, dan ulasan yang negatif ditandai dengan label -1.

Pada tahap ini, data yang telah diberi bobot seperti yang disajikan pada tabel 3.9 akan diubah ke dalam format SVM, sehingga nilai x dan nilai y didapatkan untuk menghitung vector pendukung setiap ulasan. Vektor pendukung (*support vector)* yang diperoleh akan digunakan untuk menghitung nilai a menggunakan

metode substitusi. Nilai w, dan bias (b) dihitung setelah mendapat nilai a dan akan dijadikan sebagai *hyperlane* dari model yang telah dibuat.

Sebagai contoh, dokumen S1, S2, dan S3 merupakan dokumen yang sama dengan hasil pembobotan TF-IDF pada tabel 3.9 akan diterapkan untuk mengubah data tekstual menjadi data vektor.

Tabel 3.10 Data Vector

Term	S1	S2	S3
$term_1$	0,477	0	0
$term_2$	0,477	0	0
$term_3$	0,477	0	0
$term_4$	0,176	0	0
$term_5$	0,176	0	0
$term_6$	0,176	0	0
$term_7$	0,176	0	0
$term_8$	0,477	0	0,477
$term_9$	0,477	0	0,477
$term_{10}$	0	0,352	0,176
$term_{11}$	0	0,176	0,176
$term_{12}$	0	0,352	0,176
$term_{13}$	0	0,477	0
$term_{14}$	0	0,477	0
$term_{15}$	0	0,352	0,176
$term_{16}$	0	0,477	0
$term_{17}$	0	0,477	0
$term_{18}$	0	0,477	0
$term_{19}$	0	0,477	0
$term_{20}$	0	0,477	0
$term_{21}$	0	0,477	0
$term_{22}$	0	0,477	0
$term_{23}$	0	0,477	0
$term_{24}$	0	0	0,477
Y	1	1	-1

Kernel kemudian akan dihitung menggunakan nilai x. Nilai total dari nilai x pada kolom S1 $\{term_1, term_2, \dots term_i\}$.adalah nilai untuk x_1. Begitu juga dengan nilai x_2 = S2, dan x_3 = S3. Kemudian kernelisasi akan dilakukan menggunakan fungsi kernel linear $K(x_i, x_j) = x_i . x_j^T$. Berikut ini adalah hasil perhitungan kernel linear secara manual:

$$x_1 = \begin{bmatrix} 0{,}477 & 0{,}477 & 0{,}477 & 0{,}176 & 0{,}176 & 0{,}176 & 0{,}176 \\ 0{,}477 & 0{,}477 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$K(x_i, x_j) = x_i . x_j^T = \left(\begin{bmatrix} 0{,}477 & 0{,}477 & 0{,}477 & 0{,}176 & 0{,}176 & 0{,}176 & 0{,}176 \\ 0{,}477 & 0{,}477 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} . \begin{bmatrix} 0{,}477 & 0{,}477 & 0{,}477 & 0{,}176 & 0{,}176 & 0{,}176 & 0{,}176 \\ 0{,}477 & 0{,}477 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right)$$

$K(x_i, x_j) = x_i . x_j^T$ = 1,2615.

Maka untuk nikai kernel linear selanjutnya disajikan pada tabel 3.11.

Tabel 3.11 Hasil Perhitungan Nilai X dengan kernel

$x_1 . x_1^T$	$x_1 . x_2^T$	$x_1 . x_3^T$	$x_2 . x_1^T$	$x_2 . x_2^T$	$x_2 . x_3^T$	$x_3 . x_1^T$	$x_3 . x_2^T$	$x_3 . x_3^T$
1,2615	0	0,4550	0	2,2752	0,7653	0,4550	0,7653	0,8064

Hasil perhitungan kernel yang dilakukan akan menghasilkan matriks seperti berikut:

$$x_i . x_j^T = \begin{bmatrix} x_1 . x_1^T & x_2 . x_1^T & x_3 . x_1^T \\ x_1 . x_2^T & x_2 . x_2^T & x_3 . x_2^T \\ x_1 . x_3^T & x_2 . x_3^T & x_3 . x_3^T \end{bmatrix}$$

$$x_i . x_j^T = \begin{bmatrix} 1{,}2615 & 0 & 0{,}4550 \\ 0 & 2{,}2752 & 0{,}7653 \\ 0{,}4550 & 0{,}7653 & 0{,}8064 \end{bmatrix}$$

Nilai y selanjutnya akan dihitung menggunakan cara yang sama untuk menghitung nilai x. Nilai y merupakan nilai label yang telah diberikan sebelumnya.

Tabel 3.12 Nilai Label y

y_1	y_2	y_3
1	1	-1

Kemudian perhitungan menggunakan kernel linear akan dilakukan seperti pada nilai x. Hasil perhitungan tersebut disajikan pada tabel 3.13.

Tabel 3.13 Hasil Perhitungan Nilai Y dengan kernel

$y_1.y_1^T$	$y_1.y_2^T$	$y_1.y_3^T$	$y_2.y_1^T$	$y_2.y_2^T$	$y_2.y_3^T$	$y_3.y_1^T$	$y_3.y_2^T$	$y_3.y_3^T$
1	1	-1	1	1	-1	-1	-1	1

Sehingga terbentuk matiks dari hasil perhitungan pada tabel 3.13 seperti berikut.

$$y_i.y_j^T = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$$

Kemudian setiap ulasan diubah ke dalam nilai vektor *(support vector)* = (x, y) untuk mendapat nilai ai,. Nilai x dan y untuk masing-masing dokumen diperoleh melalui persamaan berikut ini:

$$\sum_{i=1,j=1}^{n} x_i.x_j^T, (i,j = 1, \dots n) \quad (3.1)$$

$$\sum_{i=1,j=1}^{n} y_i.y_j^T, (i,j = 1, \dots n) \quad (3.2)$$

Penerapan persamaan 3.1 untuk menghitung nilai x pada setiap ulasan disajikan sebagai berikut:

$X_{S1} = x_1.x_1^T + x_1.x_2^T + x_1.x_3^T = 1{,}2615 + 0 + 0{,}4550 = 1{,}7165$

$X_{S2} = x_2.x_1^T + x_2.x_2^T + x_2.x_3^T = 0 + 0{,}2752 + 0{,}7653 = 1{,}0405$

$X_{S3} = x_3.x_1^T + x_3.x_2^T + x_3.x_3^T = 0{,}4550 + 0{,}7653 + 0{,}8064 = 2{,}0264$

Sehingga hasil akhir nilai x untuk seluruh ulasan disajikan dalam tabel 3.14.

Tabel 3.14 Nilai X Setiap Ulasan

Ulasan	S1	S2	S3
X	1,7165	1,0405	2,0264

Hasil perhitungan nilai y diperoleh dengan persamaan 3.2 yang disajikan sebagai berikut:

$$Y_{S1} = y_1.y_1^T + y_1.y_2^T + y_1.y_3^T = 1 + 1 + (-1) = 1$$

$$Y_{S2} = y_2.y_1^T + y_2.y_2^T + y_2.y_3^T = 1 + 1 + (-1) = 1$$

$$Y_{S3} = y_3.y_1^T + y_3.y_2^T + y_3.y_3^T = -1 + (-1) + 1 = -1$$

Sehingga diperoleh nilai y untuk setiap dokumen seperti yang disajikan tabel 3.15

Tabel 3.15 Nilai y pada setiap ulasan

Ulasan	S1	S2	S3
Y	1	1	−1

Nilai Y yang diperoleh selanjutnya digunakan untuk mencari *support vector*, hasil nilai akan disubstitusikan ke dalam persamaa 3.3 berikut ini:

$$\emptyset \begin{bmatrix} x \\ y \end{bmatrix} = \begin{Bmatrix} \sqrt{x_n^2 + y_n^2} > 2 \; maka \begin{bmatrix} 2 - y + (x - y) \\ 2 - x + (x - y) \end{bmatrix} \\ \sqrt{x_n^2 + y_n^2} \leq 2 \; maka \begin{bmatrix} x \\ y \end{bmatrix} \end{Bmatrix} \quad (3.3)$$

Sebagai contoh nilai x_n diperoleh dari X_{S3} dan y_n diperoleh dari Y_{S3} disubstitusikan ke dalam persamaan 3.3 seperti berikut ini:

$$\emptyset \begin{bmatrix} x \\ y \end{bmatrix} = \sqrt{2{,}0264^2 + -1^2} = 1{,}7624$$

Karena hasil yang didapatkan $1{,}7624 \leq 2$, maka:

$$\emptyset \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\emptyset \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2{,}0264 \\ -1 \end{bmatrix}$$

Hasil nilai terhadap seluruh ulasan disajikan dalam tabel 3.16

Tabel 3.16 Nilai Support Vector Setiap Ulasan

Ulasan	ØS1	ØS2	ØS3
Support Vector	$\begin{bmatrix} 1{,}7165 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1{,}0405 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 2{,}0264 \\ -1 \end{bmatrix}$

Untuk membantu menentukan nilai b yang menjadi hyperlane dan dan mendapatkan jarak tegak lurus yang optimal dengan memperhitungkan vektor positif, nilai bias 1 akan ditetapkan pada setiap *support vector* Hasilnya disajikan pada tabel 3.17 berikut ini:

Tabel 3.17 Nilai Support Vector Bias

Ulasan	ØS1	ØS2	ØS3
Support Vector bias	$\begin{bmatrix} 1{,}7165 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1{,}0405 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 2{,}0264 \\ -1 \\ 1 \end{bmatrix}$

Hasil dari nilai *support vector* selanjutnya akan dikalikan seluruhnya dengan persamaan 3.4 berikut ini:

$$\emptyset \sum_{i=1,j=1}^{n} a_i S_i^T S_j \qquad (3.4)$$

Sebagai contoh perhitungan menggunakan data yang ada pada S1 dapat diselesaikan seperti berikut ini:

$$S_i^T S_j = a_i \begin{bmatrix} 1{,}7165 \\ 1 \\ 1 \end{bmatrix}^T . \begin{bmatrix} 1{,}7165 \\ 1 \\ 1 \end{bmatrix}$$

$$S_i^T S_j = 4{,}9463 a_1$$

Hal yang sama selanjutnya diterapkan pada nilai *support vector* S2 dan S3. Setelah memperoleh seluruh hasil perhitungan, parameter a_i akan dicari menggunakan persamaan $\sum_{i=1,j=1}^{n} a_i S_i^T S_j = y_i$, sehingga memperoleh nilai a_i sebagai berikut:

$a_1 = 0{,}5557,\ a_2 = 0{,}4442,\ a_3 = 0{,}9999$

Nilai a_1 yang didapatkan kemudian digunakan untuk mencari nilai w menggunakan persamaan 2.6 pada BAB II, sehingga mengahsilkan persamaan berikut:

$$w = 0{,}5557 \,.\, 1 \,.\begin{bmatrix} 1{,}7165 \\ 1 \end{bmatrix} + \ 0{,}4442 \,.\, 1 \,.\begin{bmatrix} 1{,}0405 \\ 1 \end{bmatrix} + \ 0{,}9999 \,.\, (-1) \,.\begin{bmatrix} 2{,}0264 \\ -1 \end{bmatrix}$$

$$w = \begin{bmatrix} 0{,}9538 \\ 0{,}5557 \end{bmatrix} + \begin{bmatrix} 0{,}4621 \\ 0{,}4442 \end{bmatrix} + \begin{bmatrix} -2{,}0262 \\ 0{,}9999 \end{bmatrix}$$

$$w = \begin{bmatrix} 0{,}9538 + 0{,}4621 + (-2{,}0262) \\ 0{,}5557 + 0{,}4442 + 0{,}9999 \end{bmatrix}$$

$$w = \begin{bmatrix} 0{,}6103 \\ 1{,}9998 \end{bmatrix}$$

Perhitungan nilai b dilakukan menggunakan persamaan 2.9 dengan mensubtitusikan nilai *support vector* yang telah diperoleh. Nilai *b* diperoleh dengan menghitung nilai bias untuk setiap *support vector* dengan persamaan berikut:

$$b_i = y_i - w \,.\, x_i$$

Sehingga melalui persamaan tersebut diperoleh nilai b_i untuk masing-masing *support vector* seperti berikut:

$$b_1 = 1 - \big((0{,}6103 \,.\, 1{,}7165)\big) + (1{,}9998 \,.\, 1) = 4{,}0473$$

$$b_2 = 1 - \big((0{,}6103 \,.\, 1{,}0405)\big) + (1{,}9998 \,.\, 1) = 3{,}6348$$

$$b_3 = 1 - \big((0{,}6103 \,.\, 2{,}0264)\big) + \big(1{,}9998 \,.\, (-1)\big) = 0{,}2369$$

Nilai bias masing-masing *support vvector* kemudian akan dihitung rata-ratanya sesuai dengan persamaan 2.9, sehingga akan diperoleh perhitungan seperti berikut:

$$b = \frac{1}{N_s}(b_1 + b_2 + b_3)$$

$$b = \frac{1}{3}(4{,}0473 + 3{,}6348 + 0{,}2369)$$

$$b = 2{,}6395$$

Setelah melakukan perhitungan hingga mendapatkan nilai *hyperlane* untuk klasifikasi kelas, yaitu 2,6395, maka dapat dinilai apakah kalimat tersebut termasuk dalam kelas positif atau negatif. Dengan nilai w dan *hyperlane* tersebut, apabila hasil pengujian lebih besar dari nilai *hyperlane* maka kalimat yang diuji termasuk

dalam kelas positif dan sebaliknya apabila hasil pengujian lebih kecil dari nilai *hyperlane*, maka kalimat termasuk dalam kelas negatif.

Misalnya, data uji dilakukan untuk kalimat D1 yang memiliki nilai *support vector* (1,7165, 1). Untuk mendapat nilai kelas dari D1, dilakukan perhitungan sesuai dengan persamaan berikut:

$$w^T.\emptyset(S1) = \ [0{,}6103 \ \ 1{,}9998] \times \begin{bmatrix} 1{,}7165 \\ 1 \end{bmatrix} = 3{,}0473 > 2{,}6395$$

Hasil pengujian yang dilakukan pada D1 adalah 3,0473 yang nilainya lebih besar dibandingkan dengan nilai b yang diperoleh, yaitu 2,6395, shingga dapat disimpulkan bahwa D1 termasuk dalam kelas positif.

3.8 Desain Eksperimen

Data pada penelitian ini diperoleh dari ulasan pengguna Female Daily terhadap produk Cetaphil Gentle Skin Cleanser menggunakan *Instant Data Scrapper* yang telah disediakan oleh Google. Sekitar 1000 data yang dikumpulkan kemudian akan diberi label secara manual yang terdiri atas dua kategori, yaitu "positif" dan "negatif". Selanjutnya proses *preprocessing* dilakukan dengan tujuan memastikan data yang digunakan untuk proses selanjutnya atau pelatihan model bersih, konsisten, serta dalam format yang dapat dipahami oleh model yang telah dibuat.

Tahap awal pengujian pada penelitian ini adalah pemisahan data ke dalam data *training* dan data *testing* dengan tujuan menemukan rasio dengan performa

terbaik untuk melakukan pegujian selanjutnya. Tabel 3.18 menampilkan rasio perbandingan data yang digunakan pada penelitian ini:

Tabel 3.18 Pengujian Rasio Split Data

Pengujian ke-	Ratio Split Data	Akurasi
1	5:5	
2	6:4	
3	7:3	
4	8:2	
5	9:1	

Selanjutnya pengujian pada tahap kedua dilakukan menggunakan rasio terbaik pada pengujian sebelumnya. Pengujian pada tahap ini memiliki 2 kondisi, yaitu kondisi tanpa penggunaan *normalization* saat proses *preprocessing* dan kondisi sistem saat menerapkan *normalization* pada proses *preprocessing.* Dari hasil scenario pengujian ini, maka dapat dilihat apakah proses *normalization* pada *preprocessing* berpengaruh terhadap kinerja system atau tidak. Hasil pada skenario ini akan dimasukkan dalam tabel 3.19 berikut:

Tabel 3.19 Pengujian Penerapan Normalization

Pengujian ke-	**Normalization**	**Nilai Performa**			
		Akurasi	**Presisi**	**Recall**	**F1-Score**
1	Tidak				
2	Ya				

Pengujian selanjutnya pada penelitian ini dilakukan dengan menguji setiap rasio pada uji coba pertama dengan proporsi positif-negatif tertentu terhadap set data pelatihan. Tujuan pengujian pada kondisi ini adalah untuk mengtahui proporsi terbaik dari seluruh proporsi yang diuji dengan membandingkan nilai akurasi yang didapatkan. Hasil skenario pada pengujian ini akan dimasukkan ke dalam tabel 3.20 berikut:

Tabel 3.19 Pengujian Proporsi Positif-Negatif

Proporsi Positif-Negatif	**Rasio Split Data**			
	5:5	**6:4**	**8:2**	**9:1**
5:5				
6:4				
7:3				
8:2				
9:1				

Pengujian terakhir pada penelitian ini dilakukan dengan menggunakan beberapa fungsi kernel SVM. Tujuan pengujian pada kondisi ini adalah untuk mengtahui jenis kernel yang memiliki kinerja terbaik dengan membandingkan nilai akurasi yang didapatkan. Hasil skenario pada pengujian ini akan dimasukkan ke dalam tabel 3.20 berikut:

Tabel 3.20 Pengujian Penerapan Kernel

Nilai Hyperparameter C	Akurasi Kernel			
	Linear	RBF	Polynomial Deegree 2	Polynomial Deegree 3
0.01				
0.1				
1				
10				
100				

BAB IV

UJI COBA DAN PEMBAHASAN

Hasil evaluasi sistem yang telah dibangun dibahas pada bagian ini, beserta langkah-langkah *preprocessing*, perhitungan pembobotan kata menggunakan TF-IDF, metode *Support Vector Machine* yang digunakan untuk analisis sentimen, serta evaluasi menggunakan confusion matrix untuk mengukur nilai akurasi, presisi, recall, dan f1-score dari hasil pengujian.

4.1 Skenario Uji Coba

Skenario uji coba menjelaskan tentang beberapa skenario yang dilakukan pada system. Pengujian pertama dilakukan dengan memisahkan antara data training dan data testing. Dalam pengujian ini, perbandingan yang diterapkan terdiri dari beberapa rasio pengujian, yaitu 5:5, 6:4, 7:3, 8:2, dan 9:1. Setelah memperoleh hasil skenario yang memiliki kinerja terbaik, selanjutnya akan dilakukan pengujian terhadap kombinasi pada proses *preprocessing.* Tahap ini memiliki dua kondisi yaitu pengujian terhadap data yang menerapkan *normalization* pada proses *preprocessing* serta data yang tidak menerapkan *normalization* pada proses *preprocessing* yang bertujuan untuk melihat pengaruh penerapan proses *normalization* pada *preprocessing* teks. Pengujian pada tahap ini dilakukan menggunakan rasio dengan kinerja terbaik dari pengujian sebelumnya. Kemudian pengujian akan dilanjutkan dengan menerapkan berbagai fungsi kernel dalam SVM dengan tujuan untuk melihat fungsi kernel dengan kinerja terbaik.

4.2 Hasil Uji Coba

Data pada penelitian ini adalah data ulasan komentar terkait produk Cetaphil Gentle Skin Cleanser yang diperoleh menggunakan *tools extension* di Google Chrome, yaitu *Instant Data Scrapper*. Data yang dikumpulkan merupakan ulasan dalam rentang waktu tahun 2018 sampai dengan awal Mei 2024 berjumlah 1050 data ulasan komentar. Seluruh data kemudian dibagi ke dalam dua label secara manual, yaitu label positif dan label negatif. Dari pelabelan tersebut, diperoleh ulasan dengan label positif berjumlah 690 data, sedangkan ulasan belabel negatif berjumlah 363 data. Sebagai gambaran lebih lanjut, tabel 4.1 berikut ini menyajikan beberapa sample data dalam penelitian ini.

Tabel 4.1 Sample Dataset Penelitian

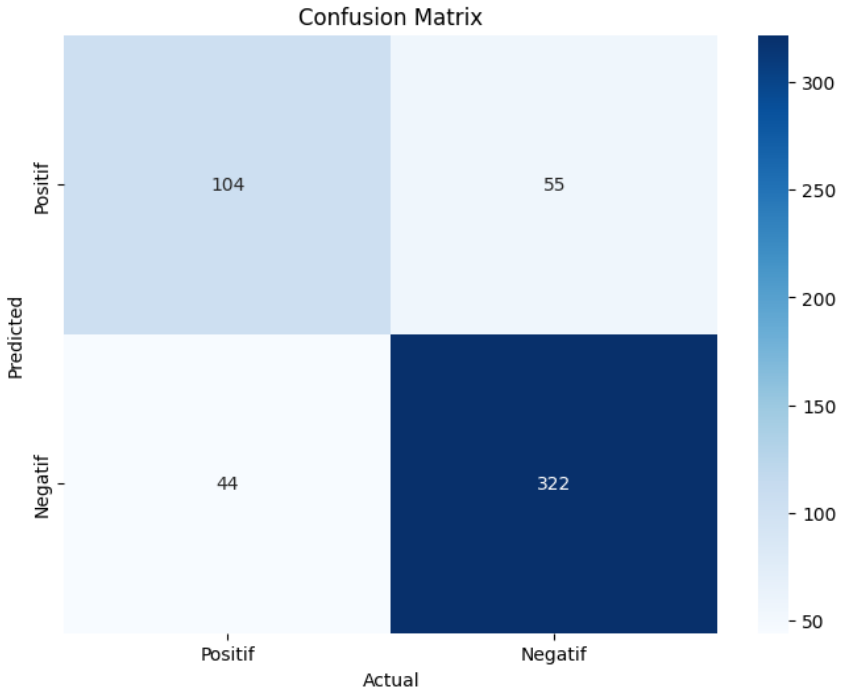
No	Ulasan	Label
1	kurang suka sm cleanser ini karena mnrt ku kurang ngangkat minyak di muka. apalagi kl abis bersihin muka pake cleanser oil based, pas abis cuci muka pk ini masih kerasa ada sisa2 minyak di muka dan kerasa kurang bersih aja. tp bener2 gentle sih dan baunya kyk hampir gaada gitu. busanya jg dikit bgt. teksturnya lumayan cair. abis pake ini muka emg jd lembab bgt sih enak, ga bikin ketarik, jd lbh kenyel gitu	negatif
2	di aku sih biasa aja, biasa bgt malah. ga ngefek sama sekali. malah aku ngerasa ky ga cuci muka gt dan ngerasa ga bersih aja. terus jerawat ku ga ngefek apa2. biasa bgt lah pokoknya. tp kalo buat dipake buat badan sih lumayan	negatif
3	kemakan review yg katanya bagus dan balik lg ke kulit kita masing2 :") dan aku lgsg beli ukuran yg paling gedenya doong karna aku pikir sebagus ituuu hahahaha... dan ternyata sama sekali ga ngaruh apa2 dimuka ku.. and its so pricey :") repurchase? of course not. karna di aku sama sekali ky gapake sabun dan gada perubahan :" preloved? cus lgsg chat aja :)	negatif
4	kemaren beli yang 125ml dan sekarang sisa sedikit, kan aku baru pake cetaphil. dan jujur, si cetaphil ini aku jadiin pelarian dari cream dokter. karena aku pengen berhenti biar gak ketergantungan gitu, eh hasilnya malah parah gilaaaa bruntusan dimana mana ? sampe aku gak berani buat lihat yg namanya kaca :(boleh saranin skincare yg lain? karena aku bener bener niat mau lepas dari cream dokter ???	negatif
5	aku pakai produk ini waktu lg hype thn 2019an. Waktu itu wajahku sedikit bruntusan, selama pemakaian di wajahku gk kasih efek buruk sama sekali, tp jg gk terlalu ngasih efek gitu sih. bruntusan masih tetep ada,, jd ini produk yg bagus cuma gk terlalu cocok sama kulitku. Sooo aku ganti deh ke produk lain. Dan produk ini lumayan awet, gak cepet habis.	positif
6	GENTLE BGT!! gak ada bau gak ada busa dan ga bikin kering samsek. aku suka banget sama dia, yang disayangkan satu, mahal doang itu aja sih huhu. overall semuanya cocok di aku, dia penyelamat bgt kalo aku mulai bruntusan, ganti fw ke cetaphil langsung sembuh beruntusanku	positif

7	Dibeliin ini sama ibuku di guardian pas muka lagi breakout cuman di aku nggak cocok dan jerawat malah pada numbuh ada matanya gitu terus pas dipake ibuku wajahnya juga gatelâ€ • padahal aku beli ori di guardian hfft mana beli yang size besar sayang banget akhirnya buat mandi aja di badan dan malah bagus kulit jadi lembap plus halus wkkw	negatif
8	Bagus banget sumpah. Ga bikin keset dan tetap lembab, tekstur nya juga bagus. No fragrance jd aman aman aja. Untuk harga agak lumayan ya, tp untuk manfaatnya si aku rasa fine. Kalo ada jerawat di punggung aku juga suka pakein jd cepat keringnya	positif
9	Aku ga cocok pake ini malah tumbuh jerawat batu dimana manaðŸ˜-terus kulanjutin eh makin parah mukaku,kujadiin sabun buat mandi eh badanku jadi ada jerawat sedih banget mana susah lagi ngilanginnya,emang ya skincare cocok cocokan yg cocok banget diorang tapi malah ga cocok diaku	negatif
10	karna waktu itu mau travelling terus liat banyak beauty vlogger yg pake ini akhirnya aku mutusin buat beli. dan dimuka aku cocok gabikin keset dan ketarik gt. abis pake ini muka jadi lembut dan kenyal banget	positif
11	WOYYYYYY INI ENAK BANGET DI MUKA. ga ketarik sama sekaliiiii ðŸ¥°ðŸ¥°ðŸ¥° tapi emang ga ada busanya sama sekali yaa wkwk. Baru pertama kali pake yg model beginian (ga ada busa) jadi kadang masih suka â€œah kurang deh kayaknya. ulang ahâ€ • ahahaha tp overall enakeun di kulit!!	positif
12	It's a good product! Untuk ukuran kecilnya sih sangat travel-friendly. Teksturnya mirip gel tapi agak cair gitu, warnanya putih, dan pas diaplikasikan enak-enak aja. Setelah dibilas gaada rasa ketarik, justru lembab. Dulu aku pakainya pas bruntusan di dahi banyak. Waktu itu setelah habis satu botol 125ml, alhamdulillah sembuh.	positif
13	saat itu pernah pake punya temen iseng nyobain, ternyata bagus gak narik kulit wajah setelah cuci muka dan masih ada efek licin licinnya setelah di pake. tergantung ke diri sendiri ya kadang ada yang gak suka kalo cuci muka itu masih licin berasa gak bersih, ada juga yang suka yang kesed	positif
14	asli ini ringan banget teksturnya, saking ringannya busanya juga dikit banget, claim nya kan emang kandungannya gak ada yang berbahaya buat ibu hamil dan ibu menyusui, jadi aku langganan pake ini sejak hamil, melahirkan, sampe sekarang menyusui.	positif
15	Gatau kenapa padahal banyak yang cocok pake ini tapi di aku ga cocok , aku pake ini pas lagi jerawatan kecil2 setelah pake ini malah jd BO , jerawatnya malah jadi gede2 , karna ak gamau ambil resiko lebih tinggi akhirnya kuputskan untuk ga make lg .	negatif

4.2.1 Hasil Uji Coba 1

Uji coba pertama dilakukan pada variasi rasio data *training* dan *testing* yang bertujuan untuk menemukan perbandingan optimal. Rasio data yang digunakan pada penelitian ini adalah 5:5, 6:4, 7:3, 8:2, dan 9:1.Melaui uji coba pada rasio 5:5 menghasilkan pembagian data yang terdiri dari 525 data untuk masing-masing data *training* dan data *testing.* Pada data *training* rasio 5:5 didapatkan jumlah data untuk

masing-masing label positif dan negatif berjumlah 322 dan 203 data. Sedangkan pada data *testing* diperoleh jumlah 366 data berlabel positif dan 159 data berlabel negatif. Hasil uji coba rasio 5:5 memperoleh tingkat akurasi sebesar 81%, presisi 70,2%, *recall* sebesar 65,4%, dan *f1-score* sebesar 68%. Gambar 4.1 berikut ini menunjukkan hasil uji coba rasio 5:5.



Gambar 4.1 Confusion Matix Rasio 5:5

Melalui hasil confusion matrix tersebut, berikut ini adalah contoh untuk menghitung nilai presisi, *recall*, dan *f1-score*. Perhitungan dilakukan menggunakan persamaan (2.11) sampai dengan persamaan (2.13) yang ada pada Bab II.

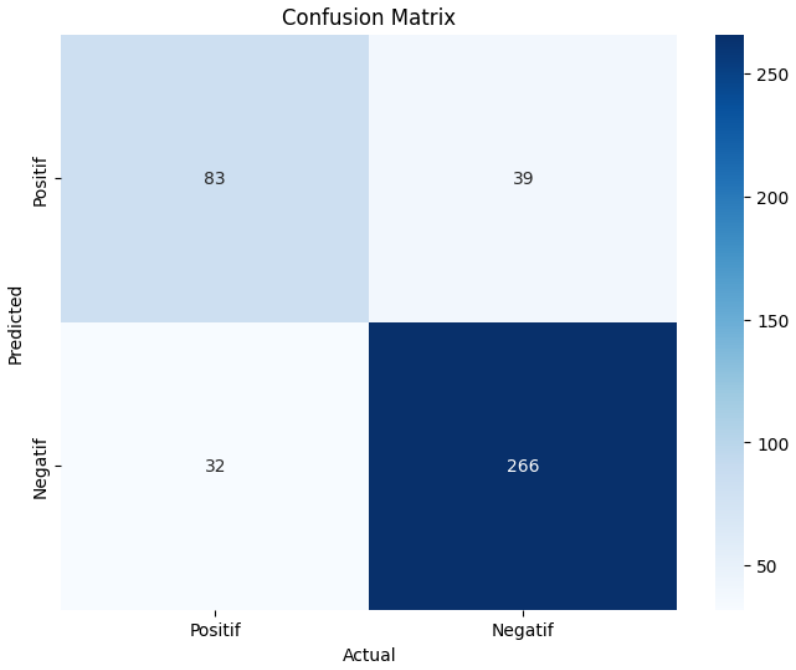
$$Recall = \frac{(104)}{104+55} \times 100\% = 65{,}4\%$$

$$Precision = \frac{(104)}{104+44} \times 100\% = 70{,}2\%$$

$$F1\ Score = \frac{2 \times 65{,}4 \times 70{,}2}{65{,}4 + 70{,}2} \times 100\% = 68\%$$

Pada pemelitian ini, pengujian dengan perbandingan data 6:4 menghasilkan pembagian data dengan 420 data uji dan 630 data latih. Pada data *training* diperoleh jumlah masing-masing data positif dan negatif adalah 390 data dan 240 data. Sedangkan dalam data *testing* yaitu 298 dan 122 data untuk masing-masing label positif dan negatif.

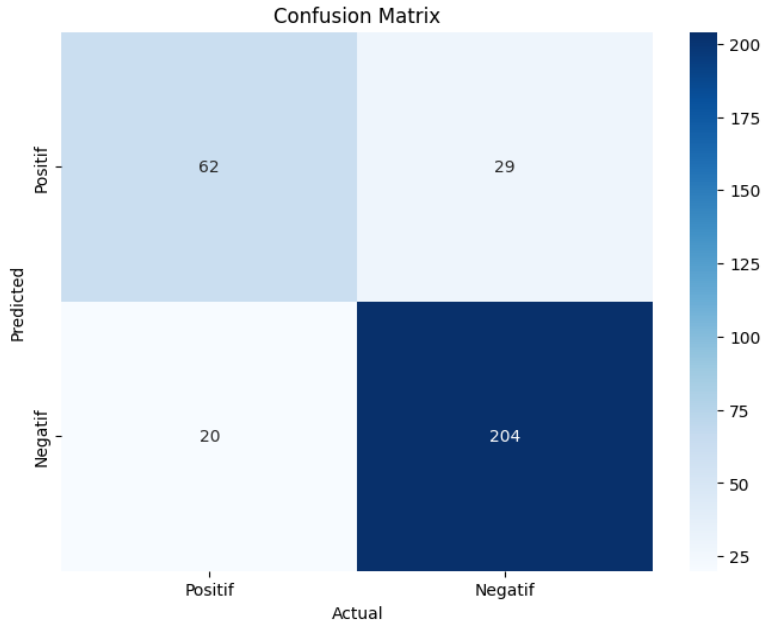
Pengujian pada tahap ini memperoleh nilai akurasi sebesar 83%, presisi 72%, *recall* 68%, dan *f1-score* 70%. Hasil uji coaba rasio 6:4 ditunjukkan pada Gambar 4.2.



Gambar 4.2 Confusion Matix Rasio 6:4

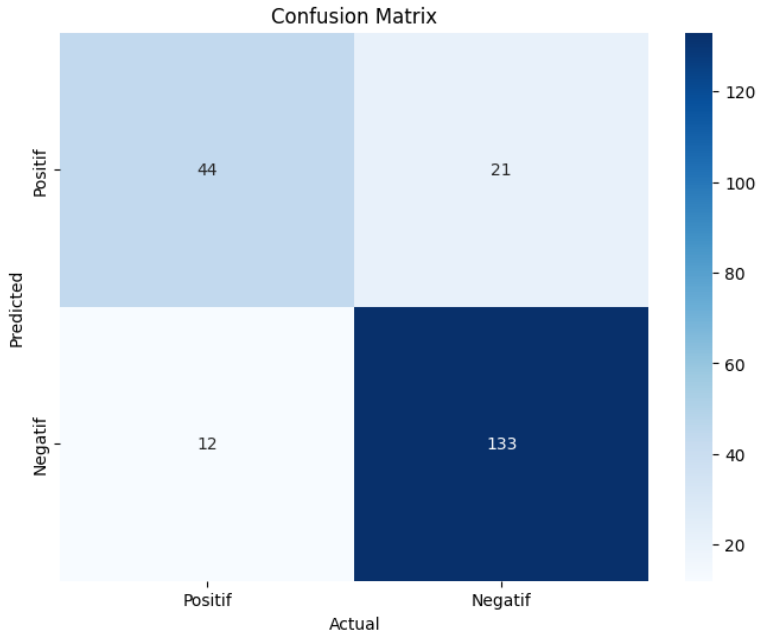
Selanjutnya dalam pengujian rasio 7:3 diperoleh jumlah data *training* yaitu 735 data dan data *testing* berjumlah 315 data. Dalam data *training*, data berlabel positif berjumlah 464 data, sedangkan yang berlabel negatif berjumlah 271 data.

Untuk data *testing*, jumlah masing-masing data dengan label positif dengan negatif adalah 224 data dan 91 data. Melalui uji coba dengan rasio 7:3 ini, diperoleh nilai akurasi sebesar 84%, presisi 76%, *recall* 68%, dan *f-1score* sebesar 72%.. hasil pengujian rasio 7:3 dapat dilihat pada gambar 4.3.



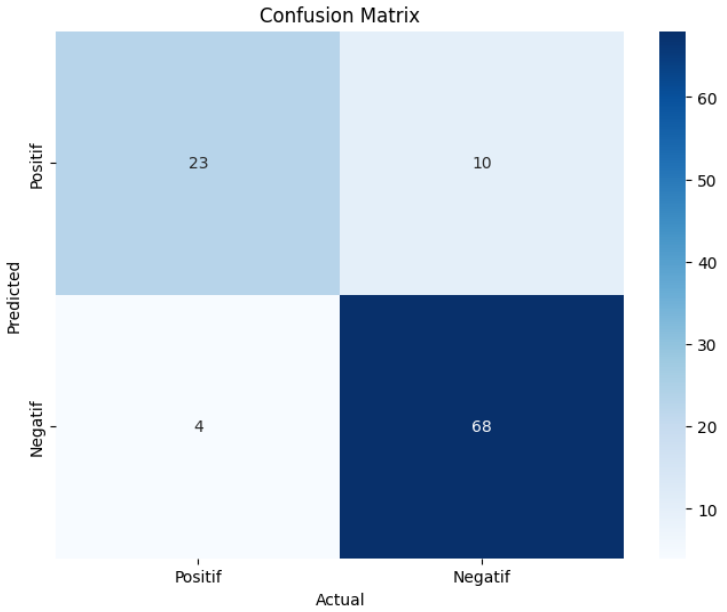
Gambar 4.3 Confusion Matix Rasio 7:3

Pengujian selanjutnya dilakukan dengan rasio 8:2. Pada pengujian ini diperoleh nilai data *training* berjumlah 840 dan data *testing* berjumlah 210. Data *training* pada pengujian ini terdapat data berlabel positif dengan jumlah 543 dan data negatif berjumlah 297. Pengujian rasio 8:2 ini menghasilkan nilai akurasi sebesar 84%, presisi 79%, *recall* 68%, dan *f1-score* 73%. Hasil pengujian menggunakan rasio 8:2 disajikan pada tabel *confusion matrix* yang terdapat pada gambar 4.4.



Gambar 4.4 Confusion Matrix Rasio 8:2

Pengujian terakhir pada uji coba rasio ini dilakukan dengan rasio perbandingan 9:1. Dalam pengujian ini dihasilkan data *training* dan *testing* masing-masing berjumlah 945 data dan 105 data. Dalam data *training*, data dengan label positif berjumlah 616, sedangkan data dengan label negatif berjumlah 329 data. Untuk data *testing*, data dengan label positif dan negatif masing-masing berjumlah 72 data dan 33 data. Melalui pengujian rasio 9:1 ini, nilai akurasi yang diperoleh adalah 87%, presisi 85%, *recall* 70%, dan *f1-score* 77%. Lebih jelasnya, hasil pengujian rasio 9:1 ditampilkan dalam gambar 4.5.



Gambar 4.5 Confusion Matix Rasio 9:1

Melalui uji coba dengan berbagai rasio data tersebut diperoleh nilai akurasi yang berbeda untuk setiap rasio. Hal ini dapat diartikan bahwa jumlah rasio data berpengaruh dalam proses uji coba sistem. Perbedaan hasil uji coba dengan variasi rasio split data disajikan pada tabel 4.2.

Tabel 4.2 Hasil Akurasi Uji Coba Variasi Rasio

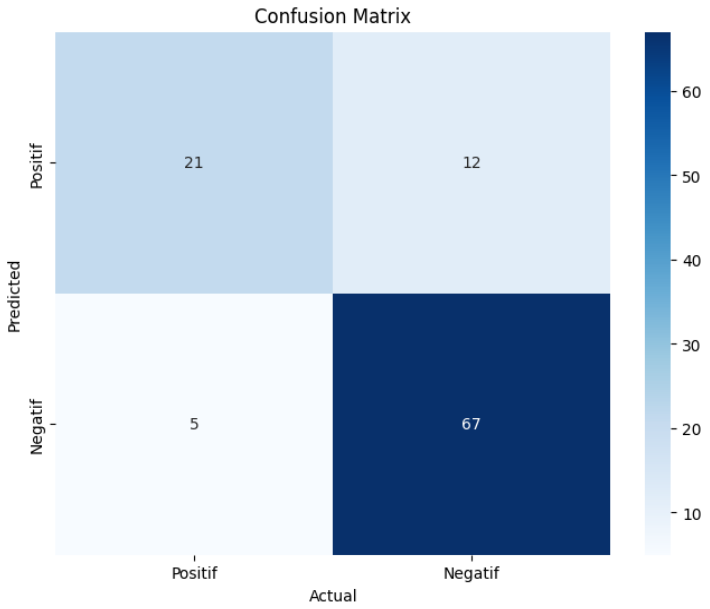
No	Rasio Split Data	Akurasi
1	5:5	81%
2	6:4	83%
3	7:3	84%
4	8:2	84%
5	9:1	86,7%

Dari tabel 4.2 dapat dilihat perbandingan hasil akurasi pada masing-masing rasio. Perbedaan hasil akurasi yang diperoleh membuktikan bahwa variasi rasio dapat berpengaruh pada hasil akurasi. Selain itu, dari uji coba yang telah dilakukan, hasil akurasi tertinggi diperoleh pada variasi rasio 9:1 dengan tingkat akurasi mencapai 86,7%. Melalui hal tersebut, maka dapat disimpulkan bahwa proses

analisis sentimen pada penelitian ini dapat bekerja optimal pada pembagian rasio 90% data *training* serta 10% data *testing*. Dari hasil pengujian ini, maka rasio 9:1 akan digunakan dalam uji coba pada tahap selanjutnya, yaitu uji coba pengaruh *normalization* pada proses *preprocessing* terhadap hasil akurasi.

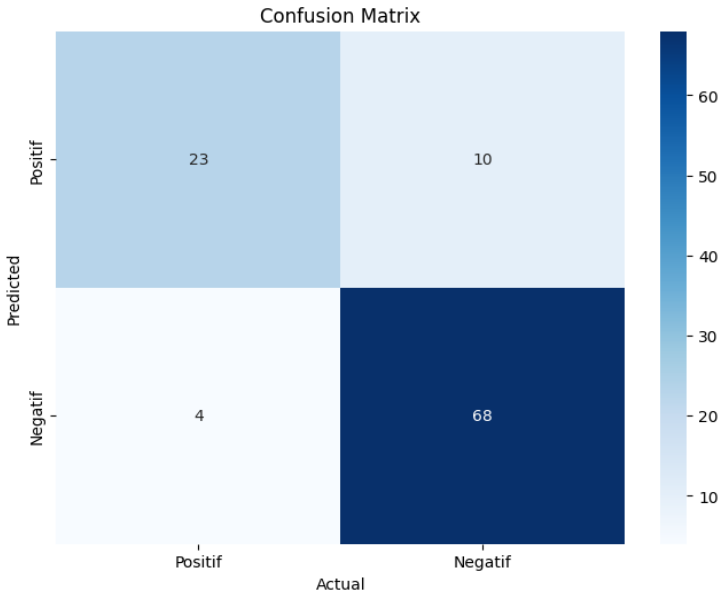
4.2.2 Hasil Uji Coba 2

Uji coba pada tahap ini dilakukan menggunakan rasio yang memiliki hasil akurasi tertinggi pada uji coba sebelumnya, yaitu rasio data 9:1. Pada tahap ini akan dibuktikan pengarus penambahan proses *normalization* pada tahap *preprocessing* terhadap performa sistem serta hasil akurasi yang didapat pada proses analisis sentimen. Gambar 4.6 menyajikan visualisasi hasil uji coba data tanpa penerapan *normalization* pada proses *preprocessing*.



Gambar 4.6 Confusion Matix Uji Coba tanpa Normalization

Berdasarkan *confusion matrix* yang disajikan pada gambar 4.6, hasil pengujian data tanpa penerapan *normalization* pada proses *preprocessing* menghasilkan nilai Akurasi sebesar 83,8%, presisi 81%, *recall* 64%, dan *f1-score* 71%. Selanjutnya sebagai perbandingan, Gambar 4.7 menyajikan visualisasi data dari hasil uji coba dengan menerapkan *normzalization* pada tahap *preprocessing.*



Gambar 4.7 Confusion Matrix Uji Coba dengan Normalization

Melalui visualisasi *confusion matrix* untuk data yang menerapkan *normalization* pada Gambar 4.7 tersebut, diperoleh nilai akurasi sebesar 86,7%, presisi 85%, *recall* 70%, dan *f1-score* 77%. Perbedaan antara hasil kedua uji coba pada tahap ini disajikan pada tabel 4.3.

Tabel 4.3 Hasil Akurasi Uji Coba Normalisasi

Pengujian ke-	Normalization	Nilai Performa			
		Akurasi	Presisi	Recall	F1-Score
1	Tidak	83,8%	81%	64%	71%
2	Ya	86,7%	85%	70%	77%

Berdasarkan hasil uji coba 2 dengan penerapan *normalization* pada tahap *preprocessing* diperoleh hasil akurasi sebesar 86,7%. Apabila dibandingkan dengan uji coba tanpa penerapan *normalization*, hasil akurasi yang diperoleh lebih rendah, yakni sebesar 83,8% dengan perbedaan sekitar 2,9%. Hal ini membuktikan bahwa penerapan *normalization* pada tahap *preprocessing* memiliki pengaruh terhadap kinerja sistem.

4.2.3 Hasil Uji Coba 3

Uji coba kali ini bertujuan untuk membandingkan pengaruh proporsi data positif dan negatif pada data train dan testing. Data pada uji coba ini menggunakan data dengan proses *preprocessing* di dalamnya. Rasio data yang digunakan sama dengan variasi rasio data pada uji coba pertama, yakni 5:5, 6:4, 7:3, 8:3, dan 9:1. Proporsi data positif dan negatif yang digunakan pada uji coba ini adalah 5:5, 6:4, 7:3, 8:2, 9:1, 2:8, 3:7, dan 4:6. Melalui percobaan tersebut didapatkan hasil seperti pada tabel 4.4 berikut.

Tabel 4.4 Hasil Akurasi Uji Coba Proporsi Positif-Negatif Setiap Rasio

Proporsi Positif-Negatif	**Rasio Split Data**				
	5:5	**6:4**	**7:3**	**8:2**	**9:1**
5:5	80%	82%%	85%	86%	87%
6:4	82%	81%	84%	83%	83%
7:3	80%	83%	83%	83%	83%
8:2	79%	79%	79%	81%	81%
9:1	72%	75%	75%	74%	74%

Berdasarkan hasil uji coba ke-3 dengan membandingkan proporsi data berlabel positif dan data berlabel negatif, dapat disimpulkan bahwa proporsi data berlabel positif dan negatif sangat berpengaruh pada proses analisis sentimen.

Melalui percobaan tersebut. rasio split data 9:1 dengan proporsi positif dan negatif 2:8 memperoleh nilai akurasi tertinggi yaitu 87%.

4.2.4 Hasil Uji Coba 4

Uji coba kali ini dilakukan untuk membandingkan fungsi kernel yang sering digunakan dalam SVM. Data pada uji coba ini menggunakan rasio data yang memperoleh akurasi tertinggi dari uji coba ke-3, yaitu rasio 9:1 dengan proporsi data berlabel positif dan negatif 5:5. Hasil uji coba perbandingan kinerja kernel disajikan pada tabel 4.5.

Tabel 4.5 Hasil Akurasi Uji Coba Kernel

Nilai Hyperparameter C	Akurasi Kernel			
	Linear	RBF	Polynomial Deegree 2	Polynomial Deegree 3
0.01	73%	75%	70%	70%
0.1	84%	73%	70%	70%
1	87%	83%	82%	76%
10	85%	85%	84%	75%
100	81%	85%	79%	83%

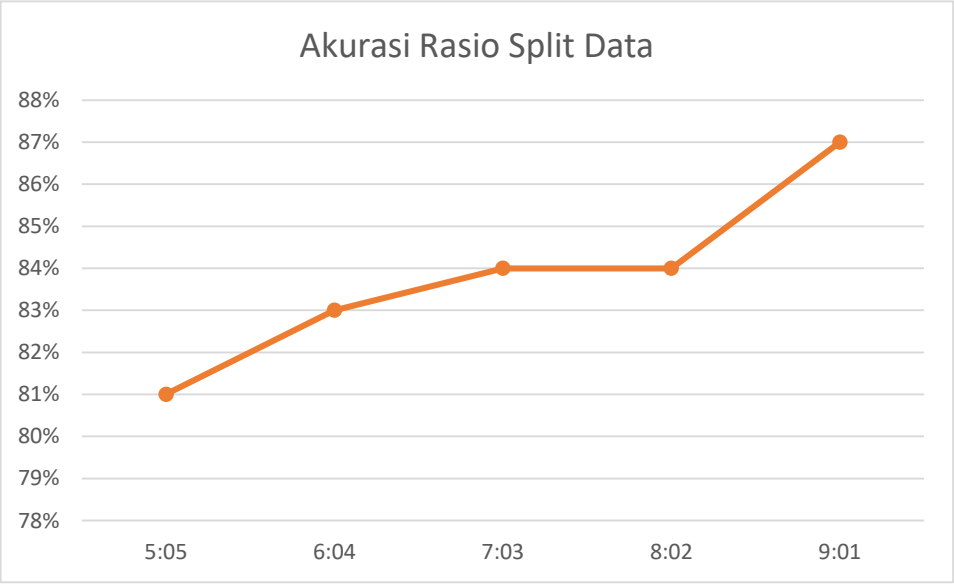
Tabel 4.5 menunjukkan hasil performa sistem dengan berbagai fungsi kernel. Berdasarkan hasil uji coba tersebut, dapat dilihat bahwa fungsi kernel berperngaruh terhadap kinerja sistem. Dengan nilai *hyperparameter* C yang diberikan pada fungsi kernel Linear, kernel Linear mencapai akurasi tertinggi saat nilai C berada pada nilai 1 dengan hasil akhir akurasi sebesar 87%. Dengan nilai *hyperparameter* C pada fungsi kernel RBF, diperoleh nilai akurasi tertinggi ketika nilai C nilai 10 dan 100, yaitu 85%. Fungsi kernel Polynomial memperoleh akurasi tertinggi yaitu 84% dengan *hperparameter* C bernilai 10. Dengan beberapa uji coba yang telah dilakukan, dapat disimpulkan bahwa konfigurasi nilai *hyperparameter* C=1 dan fungsi kernel Linear memberikan hasil terbaik untuk model SVM yang dibangun.

4.3 Pembahasan

Berdasarkan pengujian yang dilakukan terhadap 1050 data review produk dari website Female Daily,, dapat diketahui bahwa data ulasan tersebut harus diolah melalui tahap *preprocessing* yang terdiri dari beberapa proses seperti *cleaning, case folding, tokenizing, normalization, stopword removal,* hingga *stemming.* Tahap ini sangat membantu untuk sistem sehingga dapat megolah data dengan baik. Oleh karena itu, tahap ini menjadi salah satu factor penting dalam penentuan performa system yang berjalan baik atau tidak. Selanjutnya, dataset hasil *preprocessing* akan digunakan untuk melakukan uji coba yang telah dirancang. Berikut ini adalah hasil pada setiap uji coba yang dilakukan sesuai desain eksperimen yang telah dirancang.

4.3.1 Pembahasan Uji Coba 1

Berdasarkan hasil uji coba 1, didapatkan hasil performa yang berbeda dalam setiap rasio pembagian data. Gambar 4.8 menyajikan grafik perbandingan hasil akurasi setiap rasio split data.

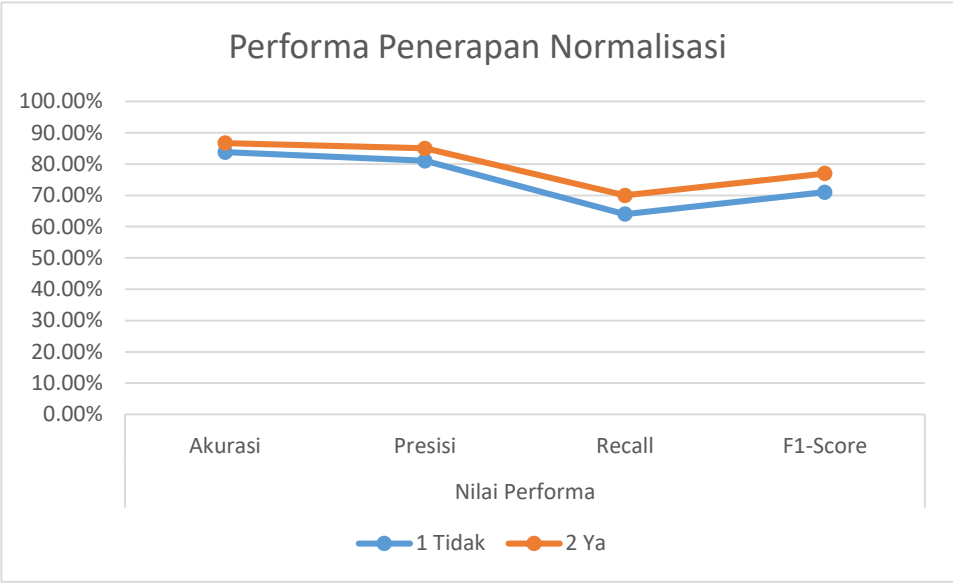


Gambar 4.8 Grafik Akurasi Uji Coba Rasio Split Data

Berdasarkan uji coba pertama didapatkan hasil, rasio split data 5:5 memiliki nilai akurasi sebesar 81%, rasio split data 6:4 memiliki nilai akurasi sebesar 83%, rasio split data 7:3 memperoleh nilai akurasi sebesar 84%, rasio split data 8:2 memperoleh nilai akurasi sebesar 84%, dan rasio dengan split data 9:1 memperoleh nilai akurasi tertinggi yaitu sebesar 86,7%. Dalam percobaan ini, terlihat bahwa nilai akurasi cenderung meningkat seiring meningkatnya jumlah data training dalam rasio split data. Hal ini menunjukkan bahwa kemampuan model untuk menggeneralisasi serta membuat prediksi akurat pada data pengujian meningkat seiring dengan jumlah data yang digunakan untuk melatih model. Dengan kata lain, jumlah data pelatihan yang lebih banyak membantu model untuk mempelajari pola dan sifat sentiment yang ada dengan lebih baik. Oleh karena itu, melalui uji coba pertama dapat disimpulkan bahwa jumlah rasio antara data training dan testing berpengaruh pada performa model.

4.3.2 Pembahasan Uji Coba 2

Selanjutnya melalui uji coba 2 didapatkan hasil uji coba dengan melakukan perbedaan fitur pada tahap preprocessing yang membandingkan penerapan *normalization* dengan tanpa menerapkannya. Tahap preprocessing pada penelitian ini menerapkan beberapa proses, yaitu *cleaning, case folding, tokenizing, stopword removal,* dan juga *stemming.* Pengujian kedua dilakukan untuk melihat pengaruh penerapan *normalization* pada proses *preprocessing* dataset. Melalui uji coba ini, performa kedua dataset hasil *preprocessing* dalam analisis sentimen akan dibandingkan. Grafik perbandingan uji coba kedua disajikan pada gambar 4.9.



Gambar 4.9 Grafik Perbandingan Nilai Performa Penerapan Normalisasi

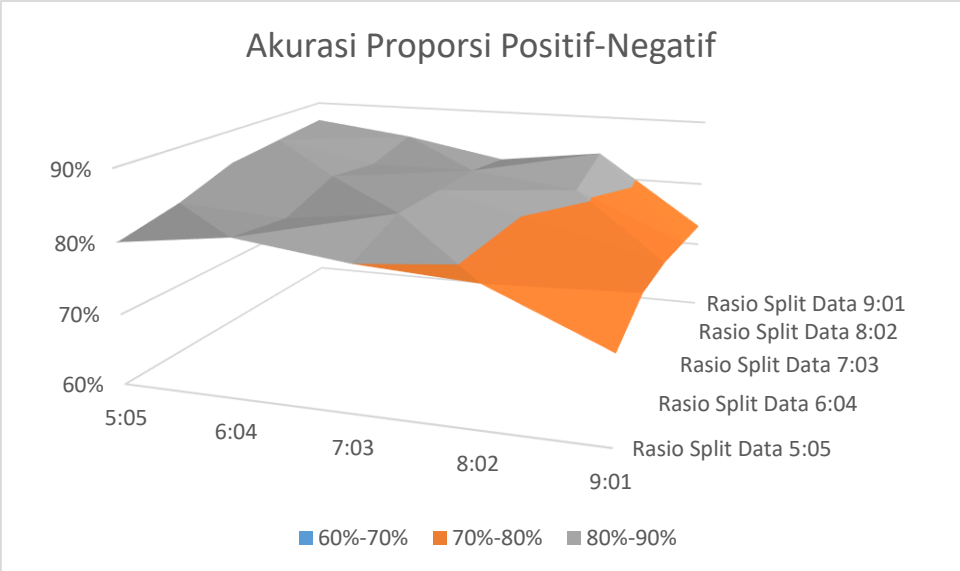
Penerapan normalisasi pada proses preprocessing memiliki pengaruh terhadap hasil akurasi uji coba 2. Berdasarkan grafik perbandingan nilai performa penerapan normalisasi yang disajikan pada gambar 4.9, dataset yang menerapkan normalisasi pada proses preprocessing cenderung memiliki nilai performa, mulai dari akurasi, presisi, *recall*, hingga *f1-score* yang lebih tinggi dibandingkan dataset yang tidak menerapkan normalisasi pada proses preprocessingnya. Hal ini terjadi karena proses normalisasi melibatkan pengubahan kata-kata pada dataset yang bertujuan untuk membuat kata-kata menjadi sama rata sehingga distribusi data menjadi lebih seragam.

Berdasarkan hasil uji coba 2, percobaan pertama menggunakan data tanpa normalisasi menghasilkan nilai akurasi sebesar 83,8% sedangkan percobaan kedua menggunakan data yang menerapkan normalisasi memperoleh nilai akurasi sebesar 87%. Akurasi pada konteks ini adalah kemampuan model SVM dalam memprediksi kelas yang ada dari seluruh total data. Artinya model SVM yang telah dibangun

melakukan prediksi lebih baik dengan data yang menerapkan normalisasi pada proses *preprocessing*. Hal ini terbukti dengan adanya peningkatan nilai akurasi sebesar 2,9% ketika dilakukan percobaan menggunakan dataset yang telah dinormalisasi. Peningkatan ini menunjukkan bahwa proses *normalization* data pada saat *preprocessing* berpengaruh untuk meningkatkan kinerja model karena memungkinkan model untuk mengidentifikasi pola serta karakteristik sentiment dengan lebih baik.

4.3.3 Pembahasan Uji Coba 3

Percobaan pada uji coba 3 dilakukan dengan tujuan untuk mengetahui pengaruh proporsi data berlabel positif dan negatif pada setiap rasio yang ada. Berdasarkan uji coba ini, diketahui bahwa baik rasio split data maupun proporsi data dalam data training memiliki pengaruh pada kinerja sistem. Grafik perbandingan antara jumlah proporsi positif dan negatif dengan rasio split data disajikan pada gambar 4.10.



Gambar 4.10 Grafik Perbandingan Hasil Akurasi Proporsi Positif Negatif

a. Akurasi Pada Setiap Rasio Split Data

Berdasarkan hasil akurasi untuk rasio split data 5:5, terlihat bahwa akurasi mulai dari 80% pada proporsi 5:5 naik menjadi 82% pada proporsi 6:4. Peningkatan ini disebabkan oleh keseimbangan yang sedikit lebih baik antara jumlah sampel dari kedua kelas, yang memungkinkan model SVM untuk lebih baik dalam memisahkan kelas positif dan negatif. Namun, terdapat penurunan ke 80% pada proporsi 7:3 dan 79% pada proporsi 8:2, yang menunjukkan bahwa keseimbangan yang kurang optimal antara proporsi positif dan negatif dapat mempengaruhi performa model. Pada proporsi 9:1, terjadi penurunan signifikan dalam akurasi menjadi 72%. Hal ini terjadi karena model yang dibangun kesulitan dalam mempelajari pola dari sampel kelas minoritas yang sangat sedikit.

Pada rasio split data 6:4, hasil akurasi menunjukkan variasi yang menarik berdasarkan proporsi positif-negatif. Pada proporsi 5:5, akurasi mencapai 82%,yang menunjukkan performa model SVM yang dibangun cukup optimal saat kedua kelas seimbang, karena model dapat belajar dari kedua kelas tanpa bias. Ketika proporsi positif-negatif ada pada 6:4, akurasi awal sebesar 82% turun menjadi 81%. Hal ini dapat terjadi karena ketidakseimbangan kecil yang membuat model lebih fokus pada kelas mayoritas (positif) dan mengurangi perhatian pada kelas minoritas (negatif). Namun, pada proporsi 7:3, akurasi meningkat menjadi 83%. Peningkatan ini dapat disebabkan oleh kemampuan model untuk dapat lebih fokus pada kelas mayoritas tanpa kehilangan terlalu banyak informasi dari kelas minoritas, sehingga model dapat mengenali pola yang lebih jelas dari kelas mayoritas. Selanjutnya, saat proporsi menjadi lebih tidak seimbang seperti pada 8:2,

nilai akurasi kembali menurun menjadi 79%. Penurunan ini disebabkan oleh model yang terlalu fokus pada kelas positif dan mulai kehilangan kemampuan untuk mengenali pola dari kelas negatif. Akurasi terus menurun secara signifikan menjadi 75% pada proporsi 9:1, dimana ketidakseimbangan yang sangat besar menyebabkan model yang dibangun akan mengalami kesulitan dalam mengenali serta memprediksi pola dari kelas minoritas, sehingga menghasilkan prediksi yang sangat bias terhadap kelas mayoritas. Terlihat sama dengan rasio split data 5:5, penurunan yang signifikan pada proporsi positif-negatif ini terjadi karena ketidakseimbangan proporsi yang meyebabkan model kesulitan untuk mempelajari pola dari kelas minoritas.

Hasil akurasi pada rasio split data pada 7:3 menunjukkan perbedaan yang signifikan berdasarkan proporsi positif-negatif. Pada saat proporsi positif-negatif 5:5, akurasi mencapai 85%, menunjukkan kinerja yang baik dari model SVM saat kedua kelas seimbang karena model dapat belajar dari kedua kelas tanpa bias. Namun, pada perbandingan proporsi 6:4, akurasi yang diperoleh turun sedikit menjadi 84% yang disebabkan oleh ketidakseimbangan kecil pada jumlah dataset positif dan negatif mulai mempengaruhi kemampuan model untuk belajar pola dari kelas minoritas. Penurunan lebih lanjut terjadi pada waktu 7:3, dengan akurasi menjadi 83%. Penurunan akurasi yang lebih besar terus terjadi pada saat proporsi positif-negatif 8:2 dan 9:1 dengan nilai akurasi 79% fsn 75% untuk msding-masing proporsi. Fenomena penurunan akurasi ini menunjukkan bahwa ketidakseimbangan proporsi yang besar menyebabkan model cenderung bias dan focus pada kelas mayoritas sehingga tidak mampu mengenali pola dengan baik.

Hasil akurasi pada rasio data 8:2 untuk proporsi 5:5 adalah 86%. Hal ini kembali menunjukkan bahwa model SVM menghasilkan kinerja yang cukup baik ketika proporsi data positif-negatif seimbang. Ketika proporsi yang berlaih ke 6:4 dan 7:3, akurasi menurun sebanyak 3%, dari 86% menjadi 83%. Penurunan nilai akurasi tersebut dapat disebabkan oleh keidakseimbangan kecil yang mulai mempengaruhi kemampuan model dalam mempelajari pola kelas minoritas. Selanjutnya pada proporsi data 8:2, akurasi kembali menurun menjadi 81%. Hal ini menunjukkan bahwa ketidakseimbangan pada proporsi yang semakin besar membuat model terlalu focus pada kelas mayoritas dan kehilangan kemampuan dalam mengenali pola untuk kelas minoritas. Penurunan yang signifikan terjadi pada proporsi data 9:1 dengan nilai akurasi 74%. Akibat ketidakseimbangan yang sangat besar ini, model SVM gagal untuk memehami pola dalam kelas minoritas sehingga menghasilkan prediksi yang sangat bias terhadap kelas mayoritas dan menurunkan akurasi secara keseluruhan. Fenomena pada rasio split data 8:2 ini menekankan pentingnya menjaga keseimbangan proporsi data antara kelas positif dan negatif untuk memastikan kinerja optimal dalam analisis sentimen. Adanya hasil akurasi yang sama pada proporsi 6:4 dan 7:3 mengindikasikan bahwa model masih mampu menggeneralisasi dengan baik meskipun terjadi ketidakseimbangan moderat. Namun, saat ketidakseimbangan menjadi sangat besar, kinerja model menurun drastis.

Pada rasio split data 9:01, hasil akurasi menunjukkan variasi yang cukup signifikan berdasarkan proporsi positif-negatif. Pada proporsi 5:5, akurasi yang diperoleh mencapai 87% yang menunjukkan performa optimal dari model SVM

saat kedua kelas seimbang. Hal ini terjadi karena model dapat belajar dari kedua kelas tanpa bias. Ketika proporsi berada pada perbabndingan 6:4, akurasi menurun menjadi 85% yang dapat terjadi karena ketidakseimbangan kecil yang mulai mempengaruhi kemampuan model untuk mempelajari polas kelas minoritas yang ada. Penurunan lebih lanjut terjadi pada proporsi positif-negatif 7:3 dengan akurasi menjadi 82%. Hal ini meunjukkan menunjukkan bahwa ketidakseimbangan yang lebih besar membuat model semakin fokus pada kelas mayoritas. Namun, pada proporsi 8:02, akurasi meningkat lagi menjadi 84%, yang dapat disebabkan oleh model yang mulai menemukan pola dalam kelas mayoritas, meskipun masih ada cukup data dari kelas minoritas untuk mempertahankan kinerja yang baik. Selanjutnya penurunan yang signifikan terjadi pada proporsi 9:01 dengan akurasi menjadi 73%. Karena ketidakseimbangan yang sangat besar, model SVM yang dibangun mengalami kesulitan dalam mengenali dan memprediksi pola dari kelas minoritas, sehingga menghasilkan prediksi yang sangat bias terhadap kelas mayoritas dan menurunkan akurasi secara keseluruhan. Adanya peningkatan akurasi pada proporsi 8:02 setelah penurunan sebelumnya menunjukkan bahwa model dapat beradaptasi dengan ketidakseimbangan tertentu, tetapi kinerja akan menurun tajam jika ketidakseimbangan menjadi terlalu besar.

b. Akurasi Berdasarkan Proporsi Positif-Negatif

Dengan proporsi 5:5, prediksi positif-negatif menunjukkan bahwa dataset terbagi secara merata antara kelas positif dan negatif, masing-masing mewakili setengah dari sampel. Dalam proporsi ini, tidak ada kelas yang dominan. Oleh karena itu, model SVM dapat belajar dari kedua kelas tanpa bias yang signifikan

terhadap satu kelas. Akurasi model meningkat secara konsisten saat jumlah data pelatihan meningkat dari 50% hingga 90%, atau dengan rasio pembagian data 5:5 hingga 9:1. Pada rasio split data 5:5, model yang dibangun mencapai akurasi 80%. Ketika data pelatihan meningkat menjadi 60% dalam rasio split data 6:4, akurasi meningkat sebanyak 2% yang menghasilkan akurasi sebesar 82%. Pada rasio split data 7:3, akurasi kembali meningkat menjadi 85%. Saat data pelatihan mecapai 80% dalam rasio split data 8:2, akurasi yang diperoleh mencapi 86%. Terakhir, pada rasio 9:1 dengan 90% data training, akurasi mencapai hasil optimal, yaitu 87%. Peningkatan ini dapat dijelaskan karena model memiliki kesempatan yang lebih besar untuk mempelajari pola yang ada di kedua kelas dengan lebih banyak data pelatihan. Dengan rasio split data yang lebih tinggi, seperti 80% atau 90% data training, model dapat memanfaatkan informasi yang lebih banyak dan lebih representatif dari populasi sebenarnya sehingga meningkatkan kemampuannya untuk mengeneralisasi dengan baik pada data baru yang tidak terlihat sebelumnya. Proporsi positif-negatif 6:04 (60% positif dan 40% negatif) mengindikasikan adanya lebih banyak sampel dari kelas positif dibandingkan kelas negatif dalam dataset. Ini berdampak pada cara model SVM mengelola dan memanfaatkan informasi dari kedua kelas dalam proses pembelajaran serta pengujian.

Pada rasio split data 5:5 dengan 50% data training dan 50% data testing, model mencapai akurasi sebesar 82%. Proporsi ini seimbang sehingga memungkinkan model untuk belajar secara adil dari kedua kelas tanpa adanya bias yang signifikan terhadap satu kelas tertentu. Kemudian, ketika data training meningkat menjadi 60% dalam rasio split data 6:4, akurasi model sedikit menurun

menjadi 81%. Penurunan ini dapat disebabkan oleh dominasi kelas positif dalam data training, yang dapat menyebabkan model kurang sensitif terhadap pola yang muncul dari kelas negatif atau minoritas saat dihadapkan pada data testing.,Namun, pada rasio split data 7:3 dengan 70% data training, akurasi kembali meningkat menjadi 84%. Kondisi ini menunjukkan bahwa model dapat memanfaatkan lebih banyak sampel dari kedua kelas untuk memperbaiki kemampuannya dalam mengenali pola-pola yang kompleks. Ketika proporsi data training mencapai 80% pada rasio split data 8:2, meskipun data testing hanya menyumbang 20%, model masih mampu mempertahankan akurasi yang kuat sebesar 83%. Ini menunjukkan ketangguhan model dalam menghadapi tantangan dari proporsi yang lebih tinggi dari data training. Terakhir, pada rasio split data 9:01 dengan 90% data training, meskipun hanya 10% dari data yang digunakan untuk testing, model berhasil mempertahankan akurasi yang tinggi sebesar 85%. Hal ini menunjukkan kemampuan model untuk menggeneralisasi pola dari kelas mayoritas dengan baik, meskipun diuji pada data testing yang sangat terbatas.

Proporsi positif-negatif 6:4 (60% positif dan 40% negatif) menunjukkan adanya lebih banyak sampel dari kelas positif dibandingkan kelas negatif dalam dataset. Ini mempengaruhi cara model SVM belajar dan menggeneralisasi pola dari kedua kelas dalam proses pembelajaran dan pengujian. Pada rasio split data 5:5 dengan 50% data training dan 50% data testing, model mencapai akurasi 82%, menunjukkan bahwa proporsi yang seimbang memungkinkan model untuk belajar secara adil dari kedua kelas tanpa bias yang signifikan. Namun, ketika data training meningkat menjadi 60% pada rasio split data 6:04, akurasi model sedikit menurun

menjadi 81%, mungkin karena dominasi kelas positif dalam data training yang mengurangi sensitivitas terhadap pola dari kelas negatif saat dihadapkan pada data testing. Pada rasio split data 7:3 dengan 70% data training, akurasi kembali meningkat menjadi 84%, menunjukkan bahwa model dapat memanfaatkan lebih banyak sampel dari kedua kelas untuk meningkatkan kemampuannya dalam mengenali pola yang kompleks. Akurasi tetap kuat pada rasio split data 8:2 dengan 80% data training (83%) menunjukkan ketangguhan model dalam menghadapi proporsi yang lebih tinggi dari data training, meskipun data testing hanya 20%. Pada rasio split data 9:1 dengan 90% data training, model berhasil mempertahankan akurasi tinggi sebesar 85%, menunjukkan kemampuannya untuk menggeneralisasi pola dari kelas mayoritas meskipun diuji pada data testing yang sangat terbatas. Secara keseluruhan, penurunan akurasi pada proporsi 6:04 mungkin disebabkan oleh kesulitan model dalam mengenali pola dari kelas minoritas (negatif) ketika proporsi data training lebih dominan dari kelas mayoritas (positif). Rasio split data 9:01 menghasilkan akurasi tertinggi karena model memiliki lebih banyak data training untuk mempelajari representasi yang lebih baik dari kedua kelas, sehingga mampu menghasilkan prediksi yang lebih akurat pada data testing yang terbatas.

Proporsi positif-negatif 7:3 (70% positif dan 30% negatif) menunjukkan distribusi lebih banyak sampel dari kelas positif dalam dataset, mempengaruhi fokus model SVM dalam mengenali pola dari kelas mayoritas selama pembelajaran. Pada rasio split data 5:5 dengan 50% data training dan testing, model mencapai akurasi 80%, menunjukkan kemampuannya belajar secara seimbang dari kedua kelas. Ketika data training meningkat menjadi 60% pada rasio split data 6:4, akurasi

naik menjadi 83%, diduga karena lebih banyak sampel positif yang memperkaya representasi kelas mayoritas dalam pembelajaran. Rasio split data 7:03 mempertahankan akurasi 83%, menunjukkan bahwa model mampu menjaga kualitas prediksi meskipun dengan proporsi yang sama antara data training dan testing. Pada rasio split data 8:2 dengan 80% data training, akurasi tetap stabil di 83%, menandakan bahwa model dapat menghadapi proporsi yang lebih tinggi dari data training tanpa mengorbankan performa. Namun, pada rasio split data 9:1 dengan 90% data training, terjadi sedikit penurunan akurasi menjadi 82%, kemungkinan disebabkan oleh fokus yang berlebihan pada kelas mayoritas dalam data training, mengurangi sensitivitas terhadap pola kelas minoritas dalam data testing yang terbatas. Keseluruhan, perubahan akurasi pada berbagai rasio split data mencerminkan bagaimana proporsi data training mempengaruhi kemampuan model SVM dalam mengenali pola dari kedua kelas, dengan stabilnya akurasi pada rasio split data 7:3 menunjukkan adaptabilitas model terhadap distribusi proporsi yang seimbang antara kelas positif dan negatif.

Proporsi positif-negatif 8:2 (80% positif dan 20% negatif) menunjukkan lebih banyaknya sampel dari kelas positif dalam dataset, yang mempengaruhi cara model SVM mengenali pola dari kelas mayoritas selama proses pembelajaran. Pada rasio split data 5:5 dengan 50% data training dan testing, model mencatat akurasi 79%, menunjukkan bahwa pembelajaran dari kedua kelas dilakukan dengan seimbang tanpa dominasi yang signifikan dari salah satu kelas tertentu. Ketika proporsi data training meningkat menjadi 60% pada rasio split data 6:4, akurasi tetap stabil di 79%, menandakan bahwa penambahan sampel dari kelas positif belum memberikan

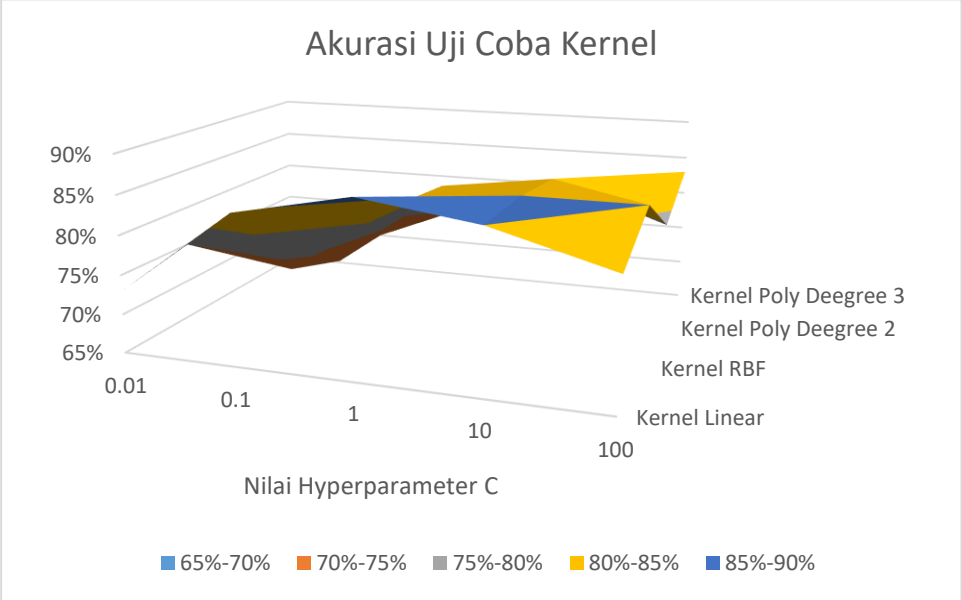
peningkatan yang signifikan dalam pengenalan pola kelas mayoritas oleh model. Pada rasio split data 7:3 dengan 70% data training, akurasi juga tetap 79%, menunjukkan konsistensi dalam performa model meskipun proporsi data training berubah. Pada rasio split data 8:2 dengan 80% data training, terjadi peningkatan akurasi menjadi 81%, kemungkinan karena fokus yang lebih besar pada kelas mayoritas dalam data training, memungkinkan model untuk lebih baik mengenali dan menggeneralisasi pola dari kelas positif. Pada rasio split data 9:1 dengan 90% data training, akurasi meningkat lebih jauh menjadi 84%, mencerminkan manfaat dari penggunaan informasi yang lebih banyak dari kelas mayoritas dalam data training untuk mengenali pola yang muncul dari data testing yang lebih terbatas. Ini menunjukkan bahwa distribusi proporsi data training secara signifikan mempengaruhi kemampuan model SVM dalam mengenali dan memproses pola dari kelas mayoritas, dengan stabilnya akurasi pada beberapa rasio split data menunjukkan kemampuan model untuk mempertahankan performa yang baik meskipun variasi dalam proporsi data training.

Berdasarkan grafik pada percobaan 3, nilai akurasi tertinggi ada pada nilai 87% yaitu ketika proporsi positif dan negatif 5:5 dengan rasio split data 9:1. Pada uji coba ini, percobaan dengan proporsi positif-negatif 5:5 cenderung menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan proporsi lainnya. Sedangkan nilai akurasi terendah diperoleh ketika proporsi positif-negatif 9:1. Hal ini mengindikasikan bahwa keseimbangan antara data positif dan negatif berpengaruh pada performa model. Hasil ini menunjukkan bahwa ketika proporsi data seimbang, model mampu belajar lebih efektif sehingga memberikan prediksi yang lebih

akurat. Lebih lanjut, uji coba ini menekankan pentingnya pemilihan proporsi data yang tepat dalam proses pelatihan model. Penggunaan proporsi 5:5 pada model memberikan model representasi lebih baik dari kedua kelas yang membantu menghindari terjadinya bias terhadap salah satu kelas. Sebaliknya, pemilihan proporsi yang tidak seimbang menyebabkan model mempelajari pola yang tidak general, sehingga menurunkan nilai akurasi. Oleh karena itu, penggunaan proporsi positif-negatif 5:5 merupakan pendekatan yang efektif untuk mencapai akurasi optimal pada penelitian ini.

4.3.4 Pembahasan Uji Coba 4

Percobaan pada bagian ini bertujuan untuk mengetahui jenis kernel yang memiliki kinerja paling optimal terhadap model yang telah dibangun. Kernel yang diuji pada percobaan ini terdiri dari tiga kernel utama, yakni kernel Linear, RBF, serta Polynomial yang diuji terhadap nilai *hyperparameter* C. Nilai C yang digunakan pada penelitian ini adalah 0,01, 0,1, 1, 10, dan 100. Pengunaan beberapa nilai parameter C dilakukan untuk menentukan nilai C dengan kinerja terbaik untuk setiap kernel. Grafik perbandingan hasil akurasi uji coba kernel disajikan pada gambar 4.10.



Gambar 4.11 Grafik Perbandingan Hasil Akurasi Uji Coba Kernel

Pada kernel Linear, hasil akurasi setiap nilai *hyperparameter* C cukup bervariasi. Pada nilai C=0,01, diperopleh akurasi sebesar 73% yang menunjukkan bahwa model tersebut terlalu sederhana atau *underfitting.* Hal ini disebabka oleh oleh toleransi kesalahan yang terlalu tinggi. Ketika nilai C ditingkatkan menjadi 0,1, akurasi meningkat secara signifikan hingga 84% yang menunjukkan bahwa model menjadi lebih ketat dalam meminimalkan kesalahan dalam data pelatihan. Akurasi puncak kernel Linear dicapai pada C=1, yaitu 87% ketika model mencapai keseimbangan optimal antara margin besar dan minimalisasi kesalahan, sehingga memungkinkan model mengenali pola yang lebih rumit. Namun, ketika angka C ditingkatkan menjadi 10, akurasinya agak menurun menjadi 85%,. Penurunan ini juga terjadi pada nilai C=100 yang menghasilkan akurasi sebesar 81%. Penurunan ini mengindikasikan bahwa model mulai mengalami overfit, terlalu berkonsentrasi pada data pelatihan, hinga kehilangan kemampuan untuk menggeneralisasi pola pada data pengujian.

Hasil akurasi pada kernel rbf menunjukkan bagaimana model SVM mengelola kompleksitas data dengan tigkat penalty kesalahan yang berbeda-beda. Nilai akurasi sebesar 75% yang diperoleh pada C=0,01 menunjukkan bahwa, walaupun memiliki marin yang cukup besar, model ini dapat menangkap pola dengan lebih baik dibandingkan kernel linear pada nilai C yang sama, namun masih menunjukkan beberapa tanda underfitting. Akurasinya turun menjadi 73% ketika nilai C dinaikkan menjadi 0,1. Penurunan ini dapat disebabkan oleh model yang terlalu ketat dalam meminimalkan kesalahan pada data pelatihan yang tidak cukup kompleks untuk mendapatkan penalti yang lebih tinggi. Akurasinya meningkat drastis menjadi 83% ketika C dinaikkan ke 1, menunjukkan bahwa model tersebut kini mencapai keseimbangan yang baik antara kompleksitas dan kemampuan generalisasi. Pada nilai C=10, akurasi yang diperoleh sebesar 85%, menunjukkan bahwa model menjaga keseimbangan antara margin yang ketat dan generalisasi yang baik, dengan sedikit peningkatan dibandingkan nilai C sebelumnya. Karena akurasi tetap konstan pada 85% pada C=100, model telah mencapai performa puncak dan tidak terpengaruh oleh peningkatan penalti kesalahan. Hal ini mungkin terjadi karena kernel RBF lebih tahan terhadap overfitting dibandingkan kernel linier pada nilai C tinggi karena fleksibilitasnya yang lebih besar dalam menangani data non-linier.

Berdasarkan nilai *hyperparameter* C, hasil akurasi pada kernel Polynomial Deegree 2 menunjukkan variasi yang cukup siginifikan. Pada nilai C=0.01, akurasi yang diperoleh sebesar 70%. Hal ini enunjukkan bahwa model mengalami underfitting yang cukup parah, dimana margin terlalu lebar dan model tidak dapat

menangkap pola kompleks dari data. Akurasinya tetap pada 70% ketika C ditingkatkan menjadi 0,1, menunjukkan bahwa meskipun dengan penalti kesalahan yang lebih tinggi, model masih belum mampu menangkap pola-pola utama secara memadai. Akurasi meningkat signifikan ke 82% ketika nilai C dinaikkan menjadi 1. Peningkatan ini menunjukkan bahwa model ini mulai memberikan keseimbangan yang lebih baik antara margin yang ketat dan kemampuan untuk menangkap pola data yang lebih kompleks. Akurasinya kembali meningkat menjadi 84% pada C=10, menunjukkan bahwa model terus memanfaatkan penalti kesalahan yang lebih tinggi untuk meningkatkan kemampuannya dalam mengenali pola dengan lebih tepat. Namun demikian, keakuratannya menurun hingga 79% ketika nilai C naik menjadi 100. Penurunan ini menunjukkan bahwa model kemungkinan mengalami overfitting, yang terjadi ketika model berupaya meminimalkan kesalahan dari set pelatihan dengan terlalu ketat, Meskipun penalti kesalahan yang sangat tinggi membantu mengurangi kesalahan dalam data pelatihan, hal ini membuat model terlalu sensitif terhadap suara dan detail kecil dalam data pelatihan, yang tidak selalu terlihat dalam data pengujian.

Berdasarkan nilai hyperparameter C, hasil akurasi pada kernel Polynomial Degree 3 menunjukkaan variasi berbeda. Jika margin terlalu besar, model tidak dapat mengekstrak pola rumit dari data, seperti yang ditunjukkan oleh akurasi sebesar 70% pada C=0,01, yang menunjukkan adanya *underfitting*, sebuah fenomena di mana margin yang terlalu besar menghalangi model untuk mengekstraksi pola rumit dari data. Serupa dengan situasi pada Polynomial Deegree 2, akurasi tetap pada 70% bahkan ketika nilai C dinaikkan menjadi 0,1,

menunjukkan bahwa meskipun penalti kesalahan ditingkatkan, model masih tidak dapat menangkap pola terkait secara akurat. Akurasi meningkat menjadi 76% ketika nilai C ditetapkan pada nilai 1. Peningkatan ini menunjukkan bahwa model mulai menyeimbangkan kebutuhan untuk mengumpulkan pola data yang semakin rumit dengan mempertahankan margin yang ketat secara lebih efektif. Pada C=10, akurasi kembali turun menjadi 75% yang menunjukkan indikasi awal bahwa model mulai mengalami *overfitting*, dimana terlalu ketat dalam meminimalisir error dari data pelatihan sehingga kehilangan kemampuan untuk menggeneralisasi pola pada data pengujian. Sekali lagi, akurasi meningkat menjadi 83% ketika nilai C dinaikkan menjadi 100. Meskipun sedikit tidak biasa , peningkatan ini dapat disebabkan oleh model yang menemukan cara baru untuk menangkap pola dalam data pelatihan yang masih relevan dengan data pengujian. Namun, hasil ini mungkin juga menunjukkan sensitivitas yang tinggi terhadap hyperparameter pada kernel polinomial yang lebih tinggi, di mana variasi kecil pada data dapat menyebabkan perubahan akurasi yang besar.

Meningkatkan nilai C pada polynomial deegree 3 pada awalnya membantu meningkatkan akurasi dengan menangkap pola yang cukup rumit dengan lebih baik jika dibandingkan dengan kernel polynomial deegree 2. Namun model polinomial derajat 3 lebih banyak menunjukkan variasi dan lebih sensitif terhadap nilai C dibandingkan dengan polynomial deegree 2. Hal ini menunjukkan bagaimana kernel Polynomial Deegree 3 lebih rumit dan kompleks untuk dioptimalkan, sehingga memerlukan pilihan nilai C yang sangat hati-hati untuk menghindari overfitting dan memastikan kemampuan generalisasi yang baik.

Dalam empat percobaan fungsi kernel yang dilakukan, dapat dilihat perbedaan hasil akhir nilai akurasi, fungsi kernel linear memperoleh nilai akurasi tertinggi 87% dengan nilai C=1, sedangkan fungsi kernel RBF memperoleh nilai akurasi tertinggi pada nilai C 10 dan 100 yaitu sebesar 85%. Fungsi kernel Polynomial memperoleh akurasi tertingginya yaitu 84% saat nilai C berada pada nilai 10 dengan degree bernilai 2. Melalui hasil uji coba kernel dapat disimpulkan bahwa pada penelitian ini fungsi kernel Linear lebih unggul dibandingkan dengan dua kernel lain, RBF dan Polynomial. Keunggulan ini dapat disebabkan oleh fakta bahwa dataset yang digunakan pada penelitian ini dapat dipisahkan secara linear. Dengan dataset yang memiliki pemisah linear yang jelas antara kelas positif dan negatif, kernel Linear mampu menangkap pola dengan lebih sederhana dan efektif.

Berdasarkan grafik pada gambar 4.11 dapat dilihat juga bahwa penggunaan nilai C yang rendah antara 0,01 dan 0,1 menghasilkan akurasi yang terkecil dibandingkan nilai C lainnya, yaitu sekitar 70% pada kernel Polynomial. Selain itu, grafik tersebut menunjukkan bahwa setiap kernel mengalami peningkatan akurasi pada nilai C=1. Dengan demikian, dalam konteks penelitian ini, penggunaan nilai parameter C =1 dapat dikatakan efektif untuk memperoleh performa model yang baik, terlebih lagi menggunakan kernel Linear. Hal ini semakin menekankan keunggulan kernel Linear yang memberikan performa lebih baik dibandingkan kernel RBF dan Polynomial.

Islam sangat menekankan pentingnya kebersihan sebagai bagian integral dari kehidupan sehari-hari dan sebagai manifestasi dari iman seseorang. Salah satu hadis yang sering dikutip dalam konteks ini adalah HR.Muslim yang berbunyi:

الطُّهُورُ شَطْرُ الْإِيمَانِ

“Kesucian itu adalah sebagian dari iman.”(HR. Muslim: 328)

Hadis ini menunjukkan bahwa menjaga kebersihan adalah bagian penting dari iman seorang Muslim. Dalam konteks modern, ini dapat diterjemahkan ke dalam praktik menjaga kebersihan tubuh dan kulit menggunakan produk-produk perawatan yang sesuai, seperti Cetaphil Gentle Skin Cleanser. Produk ini membantu membersihkan kulit wajah dari kotoran, minyak, dan debu yang dapat menyebabkan masalah kulit, sehingga berperan dalam menjaga kebersihan fisik yang juga mencerminkan kebersihan batin.

Mengacu pada review suatu produk, Cetaphil Gentle Skin Cleanser, teknik ini dimaksudkan untuk menganalisis sentimen konsumen secara efektif dan tepat.. engan mengurangi kemungkinan kesalahan dalam memahami dan menafsirkan informasi—yang seringkali mengakibatkan kesalahan dalam mengambil keputusan seperti membeli produk—cara ini dimaksudkan untuk membantu dalam menemukan data yang komprehensif dan berdasarkan fakta atau pengalaman pengguna. Penggunaan pendekatan ini diharapkan dapat memberikan kontribusi pada pengambilan keputusan yang lebih baik. Dalam surah Al-Baqarah ayat 42, Allah subhanahu wa ta'ala berfirman:

وَلَا تَلْبِسُوا الْحَقَّ بِالْبَاطِلِ وَتَكْتُمُوا الْحَقَّ وَاَنْتُمْ تَعْلَمُوْنَ

“Janganlah kamu mencampuradukkan kebenaran dengan kebatilan dan (jangan pula) kamu sembunyikan kebenaran, sedangkan kamu mengetauhuinya..”(Q.S Al-Baqarah:42)

Dalam tafsir Ibnu Katsir, ayat ini menyatakan bahwa orang Islam harus menghindari menggabungkan yang benar (haq) dengan yang salah (bathil). Ini adalah perintah bagi orang-orang beriman untuk tidak pernah goyah dari kebenaran dan jangan pernah berusaha menyembunyikannya. Oleh karena itu, prinsip ini sejalan dengan tujuan penelitian, yaitu untuk mencegah tercampurnya emosi positif dan negatif pengguna. Tujuan sistem ini tersirat dari kenyataan bahwa seperti halnya positif dan negatif, semuanya ditemukan berpasangan. Hal ini sesuai dengan surat Az-Zariyat ayat 49 yang difirmankan oleh Allah subhanahu wa ta'ala:

وَمِنْ كُلِّ شَيْءٍ خَلَقْنَا زَوْجَيْنِ لَعَلَّكُمْ تَذَكَّرُوْنَ

"Dan segala sesuatu kami ciptakan berpasang-pasangan agar kamu mengingat (kebesaran Allah)."*(Q.S. Az-Zariyat:49)*

Segala sesuatu yang diciptakan Allah berdasarkan kaidah dan standar yang telah ditetapkan dalam ajaran agama. Hal ini sesuai dengan firman Allh pada surah Al-Qamar ayat 49:

اِنَّا كُلَّ شَيْءٍ خَلَقْنٰهُ بِقَدَرٍ

"Sesungguhnya Kami menciptakan segala sesuatu sesuai dengan ukuran."(Q.S.Al-Qamar:49)

Ayat di atas memperjelas bahwa Allah SWT telah mengukur setiap ciptaan-Nya dan memberikan petunjuk kepada mereka semua. Begitu juga yang terjadi dalam kadar sentiment positif dan negatif pada analisis sentimen menggunakan metode SVM.

a. $sentimen\ positif\ \geq w.xi + b$

b. $sentimen\ negatif\ \leq w.xi + b$

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Bab sebelumnya telah menjelaskan terkait pengujian yang dilakukan untuk mengetahui efisiensi metode *Support Vector Machine* dalam analisis sentimen yang mengambil studi kasus tentang ulasan komentar produk Cetaphil Gentle Skin Cleanser di website Female Daily. Setelah melakukan beberapa pengujian, hasil akurasi terbaik dihasilkan melalui pengujian menggunakan kernel Linear pada dataset yang menerapkan normalization pada preprocessing dengan proporsi positif-negatif 5:5 pada rasio *split data* 9:1. Pengujian tersebut menghasilkan nilai akurasi sebesar 87% dengan presisi 91%, *recall* 89%, dan *F1-score* sebesar 90%. Hasil pengujian menggunakan metode SVM pada penelitian ini menunjukkan bahwa metode SVM cukup baik dalam mengklasigikasikan ulasan produk Cetaphil Gentle Skin Cleanser di website Female Daily. Skenario pengujian yang dilakukan menunjukkan bahwa kombinasi pada *preprocessing* hingga pemilihan kernel dapat berpengaruh pada hasil akhir klasifikasi saat menggunakan metode *Support Vector Machine.*

5.2 Saran

Penelitian lebih lanjut diharapkan dapat mendapatkan hasil yang lebih tepat, berdasarkan temuan penelitian penulis. Maka dari itu, untuk mendapatkan hasil terbaik, penulis menyarankan untuk melakukan hal berikut.

1. Menambahkan dataset untuk penelitian menggunakan data dari berbagai platform *e-commerce* atau sosial media lain sehingga data lebih bervariasi.
2. Melakukan percobaan berbagai kombinasi pada tahap *preprocessing* untuk menghasilkan data yang lebih bersih yang akan sangat berpengaruh pada hasil performa sistem.
3. Melakukan percobaan menggunakan algoritma berbeda selain *support vector machine* sebagai bahan perbandingan.

DAFTAR PUSTAKA

Alhaq, Z., Mustopa, A., Mulyatun, S., & Santoso, J. D. (2021). Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter. *Journal of Information System Management (JOISM)*, *3*(2), 44–49. https://doi.org/10.24076/joism.2021v3i2.558

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, *165*, 346–359. https://doi.org/10.1016/j.knosys.2018.12.005

Artama, M., Sukajaya, I. N., & Indrawan, G. (2020). Classification of official letters using TF-IDF method. *Journal of Physics: Conference Series*, *1516*(1). https://doi.org/10.1088/1742-6596/1516/1/012001

Atthahahirah, A. (2023). Pengaruh Ulasan Produk Terhadap Kepuasan Pengguna Aplikasi Female Daily. *Innovative: Journal Of Social Science Research*, *3*, 7135–7147.

Buntoro, G. A. (2017). *Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter*. *2*(1), 32–41.

Firdaus, A., & Firdaus, W. I. (2021). Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan). *Jurnal JUPITER*, *13*(1), 66.

Francia, C., & Salman. (2022). Pengaruh Electronic Word of Mouth Terhadap Minat Membeli Nature Republic. *KALBISOCIO Jurnal Bisnis Dan Komunikasi*, *9*(2), 27–36. https://doi.org/10.53008/kalbisocio.v9i2.1390

Giovani, A. P., Haryanti, T., & Kurniawati, L. (2020). *ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI*. *14*(2), 116–124.

Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *4*(2), 113. https://doi.org/10.26418/jp.v4i2.27526

Hamka, M., Alfatari, N., & Ratna Sari, D. (2022). Analisis Sentimen Produk Kecantikan Jenis Serum Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 64. https://doi.org/10.30865/json.v4i1.4740

Hendrastuty, N., Rahman Isnain, A., Yanti Rahmadhani, A., Styawati, S., Hendrastuty, N., Isnain, A. R., Rahman Isnain, A., Yanti Rahmadhani, A., Styawati, S., Hendrastuty, N., & Isnain, A. R. (2021). Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine. *Jurnal Informatika: Jurnal Pengembangan IT*, *6*(3),

150–155. http://situs.com

Jtik, J., Teknologi, J., & Samantri, M. (2024). *Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terhadap Kebijakan Pemerintah Indonesia Terkait Kenaikan Harga BBM Tahun 2022*. *8*(1), 1–9.

Karim, A. (2020). Analisis Sentimen Pada Komentar Sosial Media Instagram Layanan Kesehatan BPJS Menggunkanan Naive Bayes Classifier. *Skripsi*, *5*(3), 248–253.

Khairunnisa, S., Adiwijaya, A., & Faraby, S. Al. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *Jurnal Media Informatika Budidarma*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

Kurnianto, E., & Febriawan, D. (2023). Analisis Sentimen Perbedaan Pendapat Netizen Indonesia Terhadap Penutupan Tiktok Shop Menggunakan Algoritma Naïve Bayes. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *5*(2), 404–414. https://doi.org/10.30865/json.v5i2.7170

Luqyana, W. A. (2018). *Instagram Dengan Metode Klasifikasi Support Vector Machine*. http://repository.ub.ac.id/13396/

Madiistriyatno, H., & Alwiyah. (2023). Media Sosial dalam Manajemen Operasi dan Rantai Pasokan: Eksplorasi Masa Depan. *Jurnal MENTARI: Manajemen, Pendidikan Dan Teknologi Informasi*, *2*(1), 31–42. https://doi.org/10.33050/mentari.v2i1.372

Martini, L. K. B., & Dewi, L. K. C. (2021). Pengaruh Media Sosial Tik Tok Terhadap Prilaku Konsumtif. *Prosiding Seminar Nasional Hasil Penelitian*, *5*(1), 38–54.

Melani, B. Y., Wardhana, S. R., Hapsari, D. P., & Rozi, N. F. (2019). Analisa Kualitas Fitur Aplikasi Mobile Dengan Menggunakan Pendekatan Sentimen Grey. *Prosiding Seminar Nasional Sains Dan Teknologi Terapan*, *1*(1), 415–420.

Muktafin, E. H., Kusrini, K., & Luthfi, E. T. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Eksplora Informatika*, *10*(1), 32–42. https://doi.org/10.30864/eksplora.v10i1.390

Praghakusma, A. Z., & Charibaldi, N. (2021). Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi). *JSTIE (Jurnal Sarjana Teknik Informatika) (E-Journal)*, *9*(2), 88. https://doi.org/10.12928/jstie.v9i2.20181

Pratiwi, R. W., H, S. F., Dairoh, D., Af'idah, D. I., A, Q. R., & F, A. G. (2021). Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM). *Journal of Informatics, Information System,*

Software Engineering and Applications (INISTA), *4*(1), 40–46. https://doi.org/10.20895/inista.v4i1.387

Rahman, O. H., Abdillah, G., & Komarudin, A. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 17–23. https://doi.org/10.29207/resti.v5i1.2700

Rochmawati, N., & Wibawa, S. C. (2018). Opinion Analysis on Rohingya using Twitter Data. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012013

Taj, S., Shaikh, B. B., & Fatemah Meghji, A. (2019). Sentiment analysis of news articles: A lexicon based approach. *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, ICoMET 2019*, 1–5. https://doi.org/10.1109/ICOMET.2019.8673428

Yunita, R., & Kamayani, M. (2023). Perbandingan Algoritma SVM Dan Naïve Bayes Pada Analisis Sentimen Penghapusan Kewajiban Skripsi. *Indonesian Journal of Computer Science*, *12*(5), 2879–2890. https://doi.org/10.33022/ijcs.v12i5.3415

Zulqornain, J. A., & Adikara, P. P. (2021). *Analisis Sentimen Tanggapan Masyarakat Aplikasi Tiktok Menggunakan Metode Naïve Bayes dan Categorial Propotional Difference (CPD)*. *5*(7), 2886–2890.