Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

Perbandingan Algoritma SVM, Random Forest, KNN untuk Analisis Sentimen Terhadap Overclaim Skincare pada Media Sosial X

Ira Tri Rahmawati, Debby Alita*

Fakultas Teknik dan Ilmu Komputer, Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1]ira_rahmawati@teknokrat.ac.id, [2,*]debbyalita@teknokrat.ac.id
Email Penulis Korespondensi: debbyalita@teknokrat.ac.id
Submitted: **18/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Industri kosmetik di Indonesia, terutama produk perawatan kulit, berkembang pesat seiring dengan perubahan gaya hidup masyarakat dan kemajuan teknologi. Salah satu isu utama yang muncul adalah klaim berlebihan *overclaim*, yang dapat merugikan konsumen dan merusak reputasi perusahaan. Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma dalam analisis sentimen terhadap *overclaim skincare* di media sosial X. Algoritma yang dievaluasi meliputi *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest Neighbors* (KNN). Dataset penelitian terdiri dari 7.774 *tweet* yang dikumpulkan antara 1 Oktober hingga 30 November 2024, dengan 5.559 *tweet* setelah tahap *preprocessing*, yang terdiri dari 4.281 *tweet* sentimen negatif dan 1.275 *tweet* sentimen positif. Ketidakseimbangan data diatasi dengan menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE), dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian. Hasil penelitian menunjukkan bahwa sebelum penerapan SMOTE, algoritma *Random Forest* memiliki akurasi tertinggi sebesar 95%, diikuti oleh *Support Vector Machine* sebesar 91% dan *K-Nearest Neighbors* sebesar 80%. Setelah penerapan SMOTE, akurasi meningkat signifikan, dengan *Random Forest* mencapai 98%, *Support Vector Machine* 97%, dan *K-Nearest Neighbors* 84%. *Random Forest* terbukti menjadi algoritma terbaik, dengan kinerja tertinggi sebelum dan sesudah penerapan SMOTE, serta efektif dalam menangani kedua kelas sentimen. Penelitian ini memberikan wawasan bagi industri perawatan kulit dan regulator untuk mendeteksi dan menangani isu klaim berlebihan produk melalui pendekatan berbasis *machine learning*.

Kata Kunci: Analisis sentimen; Overclaim; Skincare; Machine Learning; SMOTE.

Abstract−The cosmetic industry in Indonesia, especially skincare products, is growing rapidly along with changes in people's lifestyles and technological advances. One of the main issues that arise is overclaiming, which can harm consumers and damage the company's reputation. This study aims to compare the performance of three algorithms in sentiment analysis of skincare overclaims on X social media. The evaluated algorithms include Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN). The research dataset consists of 7,774 tweets collected between October 1 and November 30, 2024, with 5,559 tweets after the preprocessing stage, consisting of 4,281 negative sentiment tweets and 1,275 positive sentiment tweets. Data imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), with 80% data split for training and 20% for testing. The results showed that before the application of SMOTE, the Random Forest algorithm had the highest accuracy of 95%, followed by Support Vector Machine at 91% and K-Nearest Neighbors at 80%. After the application of SMOTE, the accuracy increased significantly, with Random Forest reaching 98%, Support Vector Machine 97%, and K-Nearest Neighbors 84%. Random Forest proved to be the best algorithm, with the highest performance before and after SMOTE implementation, and was effective in handling both sentiment classes. This research provides insights for the skincare industry and regulators to detect and address product over-claiming issues through machine learning-based approaches.

Keywords: Sentiment analysis; Product Overclaim; Skincare; Machine Learnin; SMOTE.

1. PENDAHULUAN

Industri kosmetik di Indonesia berkembang pesat, didorong oleh perubahan gaya hidup masyarakat, meningkatnya penggunaan media sosial, dan tingginya jumlah penduduk usia muda. Kemajuan teknologi juga mempermudah akses informasi mengenai produk kosmetik, yang secara signifikan meningkatkan permintaan[1]. Produk kosmetik kini telah menjadi bagian tak terpisahkan dari kehidupan manusia, terutama bagi perempuan, dan telah beralih dari kebutuhan sekunder menjadi kebutuhan primer yang digunakan oleh berbagai kalangan, mulai dari bayi, anak-anak, remaja, hingga dewasa. Oleh karena itu, industri kosmetik menjadi sektor usaha yang sangat menjanjikan keuntungan besar[2]. Seiring berjalannya waktu, permintaan konsumen terhadap produk kosmetik, terutama perawatan kulit *skincare*, terus meningkat. Kehadiran perdagangan digital atau *e-commerce* memberikan kenyamanan dan kemudahan bagi konsumen untuk menemukan merek atau produk baru melalui berbagai platform online, mempermudah proses pencarian dan pembelian[3].

Fenomena ini mendorong pelaku bisnis untuk memanfaatkan media sosial sebagai sarana pemasaran interaktif, memberikan layanan, menjalin komunikasi dengan pelanggan, serta sebagai platform untuk jual beli online. Salah satu produk yang banyak dipasarkan melalui media sosial adalah skincare[4]. Tren ini memberikan peluang usaha yang besar bagi pelaku bisnis, baik kecil maupun besar. Namun, seiring dengan peluang tersebut, banyak pelaku usaha yang memanfaatkan tren ini untuk melakukan praktik kecurangan demi keuntungan yang lebih besar[5]. Industri *skincare*, yang terus tumbuh pesat, sering menggunakan strategi pemasaran agresif, termasuk *overclaim* produk. *Overclaim* mengacu pada klaim yang berlebihan mengenai manfaat dan efektivitas produk, yang sering kali tidak sesuai dengan kenyataan. Ini menjadi faktor utama yang memengaruhi keputusan konsumen dalam membeli produk[6]. *Overclaim* pada produk *skincare* dapat berbahaya bagi konsumen. Beberapa ciri *overclaim* adalah informasi yang tidak sesuai pada kemasan atau label produk, seperti ketidaklengkapan informasi mengenai

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

kandungan, dosis, garansi, khasiat, atau tanggal kedaluwarsa. Tindakan ini melanggar ketentuan yang terdapat dalam Pasal 8, 9, dan 10 Undang-Undang Nomor 8 Tahun 1999 tentang Perlindungan Konsumen[7].

Overclaim atau klaim yang berlebihan yang dilakukan oleh pelaku bisnis tentunya merugikan konsumen, karena hak-hak konsumen tidak terpenuhi. Salah satu hak konsumen adalah merasa nyaman, aman, dan terlindungi saat menggunakan produk atau layanan. Artinya, konsumen berhak untuk tidak menghadapi ketidaknyamanan, ancaman, atau risiko yang tidak diinginkan selama penggunaan produk atau layanan tersebut[8]. Dampak dari *overclaim* dapat mempengaruhi persepsi publik terhadap perusahaan. Jika klaim tersebut tidak dapat dibuktikan atau tidak terbukti benar, hal ini dapat menyebabkan kehilangan kepercayaan dari konsumen dan merusak reputasi perusahaan[9]. Fenomena *overclaim* dalam industri skincare membuat konsumen kesulitan membedakan antara produk yang benar-benar memiliki manfaat dan produk yang hanya mengandalkan klaim pemasaran yang berlebihan. Akibatnya, banyak konsumen yang tertipu dan menggunakan produk yang tidak sesuai dengan harapan, bahkan berisiko membahayakan kesehatan. Kurangnya transparansi dari pihak produsen semakin memperburuk masalah ini, seperti yang telah diungkap oleh beberapa akun media sosial, termasuk Dokter Detektif dan Dokter Richard Lee, yang menunjukkan hasil uji laboratorium yang tidak sesuai dengan klaim produk. Untuk itu, diperlukan analisis sentimen guna memahami bagaimana opini masyarakat terhadap produk-produk *skincare* yang beredar di media sosial. Selain membantu mengedukasi konsumen agar lebih selektif dalam memilih produk, analisis ini juga dapat mendorong regulasi yang lebih ketat demi melindungi kepentingan masyarakat.

Analisis sentimen memungkinkan pemetaan opini pengguna terhadap suatu produk dengan mengidentifikasi apakah sentimen yang muncul cenderung positif atau negatif. Data yang diperoleh dari analisis ini dapat menjadi alat penting bagi perusahaan dalam menentukan strategi pemasaran dan pengembangan produk. Di sisi lain, informasi ini juga bermanfaat bagi konsumen yang ingin mengetahui reputasi suatu produk sebelum menggunakannya[10]. Dengan memanfaatkan data dari media sosial, analisis sentimen dapat mengklasifikasikan opini, ulasan, atau pendapat berdasarkan emosi yang diungkapkan terkait suatu topik tertentu[11]. Hal ini menjadikannya metode yang sangat efektif dalam mengukur persepsi publik secara luas, memberikan wawasan mendalam mengenai opini masyarakat, serta membantu pengambilan keputusan yang lebih tepat dan strategis, baik oleh produsen dalam merancang serta memasarkan produk, maupun oleh konsumen dalam memilih produk yang aman dan sesuai dengan kebutuhannya.

Analisis sentimen telah menjadi bidang penting dalam memahami opini publik di berbagai domain. Beragam algoritma seperti SVM, KNN, dan *Random Forest* telah dibandingkan untuk menemukan metode terbaik dalam berbagai konteks. Kajian terhadap penelitian terdahulu memberikan gambaran performa algoritma ini dalam skenario yang berbeda seperti penelitian yang dilkukan oleh. Wicaksono dkk (2023) membandingkan algoritma *Random Forest*, SVM, dan KNN untuk analisis sentimen berbasis aspek pada review *Female Daily*, dengan SVM (*kernel linear*) mencapai akurasi terbaik sebesar 67,10%[12]. Wardani dkk (2022) mengkaji analisis sentimen kegiatan trading di *Twitter*, menunjukkan KNN memiliki akurasi tertinggi 0,999, lebih unggul dibandingkan *Random Forest* 0,994 dan SVM 0,992[13]. Muttaqin dan Kharisudin (2021) menemukan bahwa SVM (*kernel linear*, C=1) unggul dengan akurasi 87,98% dalam analisis sentimen aplikasi Gojek, dibandingkan KNN yang memperoleh 82,14%[14]. Sari dan Suryono (2024) menunjukkan akurasi hampir setara antara *Random Forest* 91% dan SVM 90% untuk analisis sentimen *metaverse* [15]. Yanti dkk (2023) membandingkan KNN 72,86% dan *Random Forest* 73,37% untuk analisis sentimen isu minyak goreng di Indonesia[16]. Septiana dan Alita (2024) menemukan bahwa SVM 80% lebih akurat dibandingkan *Random Forest* 78% dalam analisis sentimen *quick count* Pemilu 2024[17].

Berdasarkan penelitian sebelumnya, terdapat kesenjangan dalam analisis sentimen yang spesifik terhadap isu *overclaim skincare* pada media sosial X. Meski algoritma SVM, *Random Forest*, dan KNN telah banyak digunakan dalam berbagai kasus, belum ada penelitian yang secara khusus membahas perbandingan performa algoritma ini pada *overclaim skincare*. Faktor seperti karakteristik data opini pada media sosial, pola bahasa yang digunakan, serta relevansi sentimen terhadap *overclaim* memerlukan pendekatan analisis yang lebih spesifik. Oleh karena itu, penelitian ini bertujuan membandingkan performa algoritma SVM, *Random Forest*, dan KNN dalam analisis sentimen *overclaim skincare* di media sosial X, menentukan algoritma terbaik, memberikan rekomendasi bagi industri dan regulator, serta menyumbang kajian akademik terkait penggunaan *machine learning* untuk isu spesifik. Hasilnya diharapkan menjadi referensi studi lanjutan dan solusi berbasis data bagi industri *skincare.*

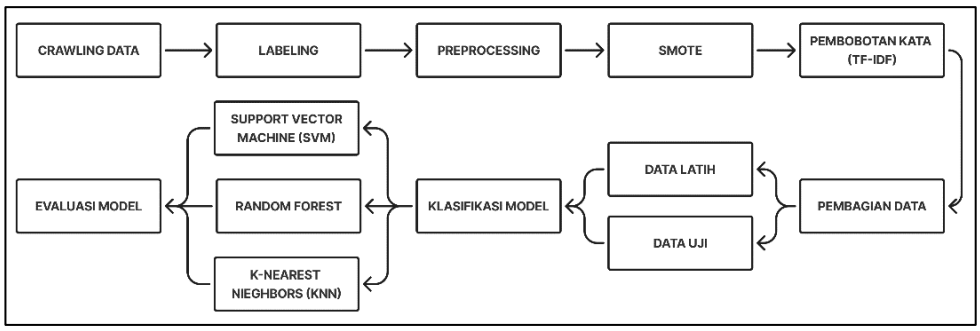
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini bertujuan untuk menganalisis sentimen terkait *Overclaim Skincare* di Media Sosial X melalui tahapan sistematis. Berdasarkan gambar 1, penelitian ini dimulai dengan *crawling* data dari media sosial X, di mana data dilabeli berdasarkan kategori atau sentimen seperti positif dan negatif. Data kemudian diproses melalui tahapan *preprocessing* untuk membersihkan dan mempersiapkannya, termasuk *cleansing*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Untuk mengatasi ketidakseimbangan data, digunakan teknik SMOTE *Synthetic Minority Oversampling Technique*, yang menciptakan data sintetis pada kelas minoritas. Data diberi bobot menggunakan metode TF-IDF *Term Frequency-Inverse Document Frequency*, lalu dibagi menjadi data latih dan data uji. Proses klasifikasi dilakukan menggunakan algoritma *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

Neighbors (KNN). Tahap akhir adalah evaluasi model dengan metrik seperti akurasi, *precision*, *recall*, dan *F1-score* untuk menentukan model terbaik.Langkah-langkah penelitian ini disajikan dalam Gambar 1.



Gambar 1. Metode Penelitian

2.2 Pengumpulan Data

Data dikumpulkan dari media sosial X melalui proses crawling. Metode *crawling* adalah teknik otomatis untuk mengumpulkan data dari situs web secara terorganisir dan efisien, sehingga cocok digunakan untuk penelitian, analisis, atau pengembangan aplikasi[18]. Pengumpulan data dilakukan menggunakan *Python* dengan bantuan *library Harvest* serta *Node.js*, yang dioperasikan melalui *code editor Google Colab*[19]. Proses *crawling* dapat dilihat pada Gambar 2.

Gambar 1. Proses *Crawling* Data

Berdasarkan gambar 1, data diakses menggunakan Token API untuk autentikasi, kemudian tweet dikumpulkan melalui proses *harvesting*. Data tersebut diproses di *Google Colaboratory* menggunakan *skrip Python*, sebelum akhirnya disimpan dalam format *Excel*.data yang digunakan dalam penelitian ini adalah *tweet* berbahasa Indonesia dari media sosial X, dengan kata kunci "*Overclaim Skincare*" dan "*Overclaim Doktif*." Sebanyak 7.774 *tweet* berhasil dikumpulkan, mencakup periode antara 1 Oktober 2024 hingga 30 November 2024.

2.3 Labeling

Proses pelabelan dilakukan secara otomatis menggunakan *textblob*, yang mengklasifikasikan *tweet* berdasarkan kelas opini, yaitu positif dan negatif. *TextBlob* menghitung nilai *polarity* dan *subjectivity* untuk menentukan kategori teks. *Polarity* menggambarkan emosi dalam teks, sementara *subjectivity* menunjukkan jenis teks tersebut[20]. Teks dengan *subjectivity* tinggi dianggap sebagai opini, sementara teks dengan *polarity* tinggi dianggap positif. Berdasarkan nilai *polarity*, teks dikategorikan sebagai positif dan negatif[21].

2.4 Preprocessing

Preprocessing adalah proses mempersiapkan data agar siap diolah, yang merupakan salah satu langkah terpenting dalam *text mining*. Tahapan *preprocessing* meliputi beberapa langkah penting untuk membersihkan dan mengatur data sebelum analis[22]. Penelitian ini menggunakan lima tahapan *Preprocessing* yaitu *Cleansing*, *Case Folding*, *Tokenizing*, *Stopword Removal*, dan *Stemming*, dijelaskan sebagai berikut:

2.4.1 Cleansing

Cleansing merupakan tahap menghapus karakter *non-alfabet* untuk mengurangi noise dalam data. Karakter yang dihilangkan meliputi tanda baca seperti titik (.), koma (,), tanda tanya (?), dan tanda seru (!), serta simbol-simbol seperti '@' untuk *username*, '#' untuk *hashtag*, emotikon, dan tautan *website* atau URL[23].

2.4.2 Case Folding

Case folding adalah proses mengubah semua huruf kapital dalam dokumen menjadi huruf kecil untuk meningkatkan konsistensi dan akurasi analisis teks, serta mencegah perbedaan penulisan yang tidak diperlukan[24].

2.4.3 Tokenizing

Tokenizing adalah proses seleksi pemotongan kata dalam sebuah kalimat dengan menggunakan pemisah seperti koma (,), titik (.), dan tanda pemisah lainnya[25]. memecah teks menjadi unit-unit yang lebih kecil, seperti kata atau frasa yang disebut token[26].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

2.4.4 Stopword Removal

Stopword adalah tahap yang berfungsi untuk menghapus kata-kata umum yang sering muncul namun kurang relevan dengan konteks teks. Kata-kata yang dihilangkan ditentukan berdasarkan daftar *stopword* yang telah ditetapkan sebelumnya[27].

2.4.5 Stemming

Stemming adalah proses menghilangkan imbuhan pada awal dan akhir kata untuk mendapatkan bentuk kata dasarnya. Proses ini dilakukan menggunakan pustaka *Sastrawi* pada bahasa pemrograman *Python*[28].

2.5 SMOTE

SMOTE (*Synthetic Minority Oversampling Technique*), adalah metode yang sering digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset. Metode ini membuat sampel baru dari kelas *minoritas* dengan menyintesis *instance* baru[29] SMOTE berfungsi dengan menambah jumlah data pada kelas *minoritas* secara buatan. Proses ini dilakukan dengan menghasilkan data sintetis baru yang berdasarkan pada data yang sudah ada di kelas *minoritas*. Data *sintetis* tersebut dibuat dengan mempertimbangkan karakteristik dari tetangga terdekat data asli[30]. Tujuannya adalah untuk menyeimbangkan jumlah data antar kelas sehingga dapat meningkatkan kinerja model klasifikasi.

2.6 Pembobotan Kata TF-IDF

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah proses yang mengubah data tekstual menjadi data numerik untuk memberikan bobot pada setiap kata atau fitur. TF-IDF merupakan ukuran statistik yang digunakan untuk menentukan seberapa penting sebuah kata dalam sebuah dokumen. TF (*Term Frequency*) mengukur seberapa sering kata muncul dalam sebuah dokumen, yang menunjukkan seberapa relevan kata tersebut dalam konteks dokumen itu. DF (*Document Frequency*) mengukur seberapa banyak dokumen yang mengandung kata tersebut, yang menggambarkan seberapa umum kata itu. IDF (*Inverse Document Frequency*) adalah kebalikan dari nilai DF. Pembobotan kata dengan TF-IDF dihitung dengan mengalikan nilai TF dengan IDF[31]. Bobot kata akan semakin besar jika kata tersebut sering muncul dalam suatu dokumen, dan semakin kecil jika kata tersebut muncul di banyak dokumen. Berikut adalah rumus untuk menghitung TF-IDF (*Term Frequency-Inverse Document Frequency*). Rumus TF-IDF disajikan dalam persamaan 1 berikut.

$$TF * IDF(d,t) = TF(d,t) * log\frac{N}{df(t)} \quad (1)$$

TF merupakan nilai dari *term frequency*, sedangkan IDF(d,t) adalah nilai dari *inverse document frequency*, yang mengukur pentingnya sebuah term t dalam dokumen d. Dalam hal ini, (d,t) menunjukkan banyaknya term t pada dokumen d, dan N merujuk pada jumlah total dokumen dalam kumpulan dokumen. Selain itu, df(t) adalah jumlah dokumen yang mengandung term t, yang digunakan untuk menghitung nilai IDF.

2.7 Pembagian Data Latih dan Data Uji

Pada tahap pembentukan data, dataset dibagi menjadi dua bagian, yaitu Data dibagi menjadi 80% untuk pelatihan *training* dan 20% untuk pengujian *testing*. Data latih digunakan untuk melatih algoritma klasifikasi, sedangkan data uji digunakan untuk mengevaluasi performa dan akurasi model yang telah dilatih[32]. Setelah pembagian data, langkah berikutnya adalah menerapkan algoritma klasifikasi seperti SVM, *Random Forest*, atau KNN untuk menemukan pola dalam data latih, agar model dapat mengklasifikasikan data uji dengan akurasi yang optimal.

2.8 Klasifikasi SVM, Random Forest, KNN

2.8.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang efektif untuk klasifikasi, *regresi*, dan prediksi[33]. SVM bekerja dengan mencari *hyperplane* terbaik untuk memisahkan dua kelas, memaksimalkan *margin* antara keduanya. *Hyperplane* ini dipilih agar jarak terdekat ke setiap kelas, yang disebut *margin*, menjadi maksimum. Titik data yang berada dekat margin ini disebut *vektor* pendukung, yang menentukan posisi *hyperplane*[34]. Rumus perhitungannya adalah sebagai berikut dapat dilihat pada persamaan 2,3 dan 4.

$$\{(x_i, y_i)\}_{i=1}^{N} \quad (2)$$

Pasangan data (x_i, y_i) terdiri dari x_i vektor fitur yang merepresentasikan atribut data ke-i dan y_i label kelas yang menentukan kategori atau nilai target, seperti +1 atau −1 dalam klasifikasi biner, Terdapat N total pasangan dalam dataset.

$$w = \sum_{i=1}^{N} a_i, y_i, x_i \quad b = -\frac{1}{2}(w.x^{+} + w.x^{-}) \quad (3)$$

Menghitung nilai w dan b, bobot w menentukan arah hyperplane, sedangkan bias b menggesernya untuk menyesuaikan dengan data. Kombinasi keduanya membentuk persamaan $w \cdot x + b = 0$, yang memisahkan kelas data.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

$$f(x) = w.x + b \quad atau \quad f(x) = \sum_{i=1}^{m} a_1, y_i K(x, x_i) + b \tag{4}$$

Fungsi Keputusan klasifikasi *sign* (f(x)), dimana $f(x)$ Hasil prediksi kelas dari data masukan x, $K(x, x_i)$ Fungsi *kernel* untuk menghitung kesamaan antara x dan x_i, m Jumlah data pelatihan yang relevan.

2.8.2 Random Forest

Random Forest adalah salah satu metode *machine learning* yang digunakan untuk klasifikasi data dalam jumlah besar dan merupakan pengembangan dari metode *Classification* and *Regression Tree*[35]. *Random Forest* merupakan gabungan dari berbagai teknik pohon keputusan yang dikombinasikan ke dalam satu model. Ada tiga langkah utama dalam metode *Random Forest*, yaitu, melakukan *bootstrap* sampling untuk membangun pohon keputusan, setiap pohon keputusan menggunakan *prediktor* yang dipilih secara acak, *Random Forest* menghasilkan prediksi dengan menggabungkan hasil dari setiap pohon keputusan melalui metode *voting mayoritas* untuk klasifikasi, atau dengan menghitung rata-rata untuk *regresi*[36]. Rumus perhitungannya adalah sebagai berikut pada persamaan 6.

$$f(x) = Average(f_1(x), f_2(x), \dots, f_n(x)) \tag{6}$$

Dalam algoritma *ensemble* seperti *Random Forest*, $f(x)$ adalah hasil prediksi akhir model terhadap input data (x), yang diperoleh dengan menggabungkan hasil prediksi dari beberapa pohon keputusan. Setiap pohon keputusan ke-n menghasilkan prediksi $f_{1-n}(x)$, dan prediksi akhir $f(x)$ dihitung dengan menggabungkan hasil dari semua pohon, menggunakan *voting* untuk klasifikasi atau rata-rata untuk *regresi*. Dengan cara ini, *Random Forest* meningkatkan akurasi prediksi dengan memanfaatkan banyak pohon keputusan.

2.8.3 K-Nearest Neighbors (KNN)

Algoritma K-Nearest Neighbors (K-NN) merupakan metode klasifikasi yang tidak membuat representasi kategori secara eksplisit, melainkan bergantung pada label kategori yang ada pada dokumen pelatihan yang serupa dengan dokumen uji. K-NN bekerja dengan cara mengklasifikasikan objek berdasarkan data pelatihan yang memiliki jarak terdekat. Prinsip dasar dari algoritma ini adalah menggunakan jarak terpendek antara sampel uji dan sampel pelatihan untuk menentukan kelas objek tersebut[37]. Rumus perhitungannya adalah sebagai berikut pada persamaan 7.

$$d(x, y) = \sqrt{\sum_{i}^{n} = 1 (x - y)^2} \tag{7}$$

Jarak *euclidean* $d(x, y)$ mengukur jarak antara dua titik data x dan y dalam ruang fitur, dengan x_i dan y_i sebagai fitur ke-i dari masing-masing titik, dan nnn adalah jumlah fitur. Jarak dihitung dengan mengkuadratkan selisih fitur, menjumlahkannya, lalu mengambil akar kuadrat dari hasilnya.

2.9 Evaluasi

Evaluasi model dilakukan dengan menggunakan metrik seperti *accuracy*, *precision*, *recall*, dan *F1 score*. Akurasi menunjukkan seberapa baik model dalam mengklasifikasikan data dengan benar. Metrik ini dihitung dengan membagi jumlah prediksi yang tepat dengan total jumlah prediksi yang dilakukan[38]. Hasil evaluasi ini dianalisis lebih lanjut untuk mengidentifikasi kelebihan, kekurangan, serta potensi perbaikan bagi model[39]. Rumus perhitungannya adalah sebagai berikut disajikan pada persamaan 8,9,10 dan 11.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{8}$$

$$precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$f1 - score = \frac{2xPresisixRecall}{PresisixRecall} \tag{11}$$

$True\ Positive\ (TP)$ Jumlah data yang berhasil dikenali dengan benar sebagai positif oleh model, $False\ Positive\ (FP)$ Jumlah data yang salah diklasifikasikan sebagai positif oleh model, $True\ Negative\ (TN)$ Jumlah data yang berhasil dikenali dengan benar sebagai negatif oleh model, $False\ Negative\ (FN)$ Jumlah data yang salah diklasifikasikan sebagai negatif oleh model.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada tahap pengumpulan data, dilakukan proses *crawling tweet* menggunakan metode *tweet harvest* di media sosial X. Data yang berhasil dikumpulkan berjumlah 7.774 *tweet*, dengan rentang waktu pengumpulan mulai dari 1 Oktober hingga 30 November 2024. Proses ini dilakukan dengan teknik *crawling* menggunakan kata kunci tertentu, salah satunya adalah "*Overclaim Skincare*" dan "*Overclaim* Doktif".Data yang telah diperoleh kemudian disimpan dalam

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

format *excel* untuk digunakan dalam analisis sentimen. Pengumpulan data dilakukan dengan memanfaatkan *twitter* API *(Application Programming Interface)*, yang diakses melalui *Google Colab* menggunakan bahasa pemrograman *Python*. Hasil data yang diperoleh dari proses *crawling* ini ditampilkan pada Tabel 1.

Tabel 1. Hasil Pengumpulan Data *Tweet*

No	Username	Tweet
1	dinsskuyyy	Guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim https://t.co/MlVbRoZBvM
7774	Rafitheda	bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang https://t.co/JL2yraE3ub

Berdasarkan sampel data pada Tabel 1 yang berhasil dikumpulkan, langkah berikutnya adalah melakukan proses pelabelan. Proses ini bertujuan untuk mengidentifikasi dan menentukan apakah setiap komentar memiliki sentimen positif atau sentimen negatif.

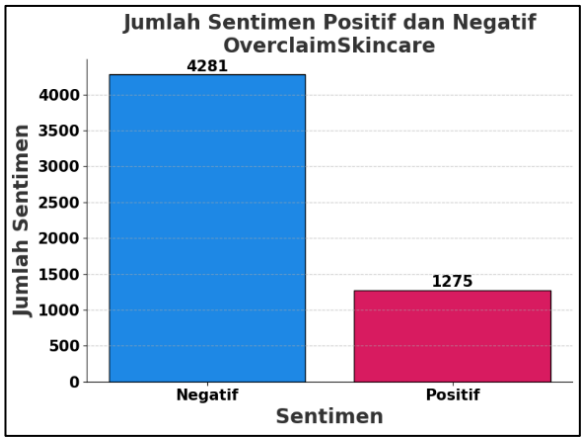
3.2 Pelebelan Data

Tahap berikutnya adalah proses *labeling*, yaitu pemberian label positif dan negatif pada data berdasarkan kriteria yang telah ditentukan. Langkah ini bertujuan untuk mengelompokkan data sesuai dengan sentimen yang terkandung dalam teks, sehingga analisis sentimen dapat dilakukan secara lebih tepat dan akurat. Hasil dari proses ini ditampilkan pada Tabel 2.

Tabel 2. Hasil Pelabelan

No	Username	Tweet	Sentiment_Score	Labeling
1	dinsskuyyy	Guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim https://t.co/MlVbRoZBvM	0.25	Positif
5559	Rafitheda	bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang https://t.co/JL2yraE3ub	-0.4	Negatif

Tabel 2 menunjukkan proses pemberian label sentimen *tweet* menggunakan *TextBlob*, dengan perhitungan berdasarkan nilai *polarity*. Jika nilai *polarity* lebih dari 0, *tweet* diberi label Positif, sedangkan jika nilainya kurang dari 0, diberi label Negatif. Sebagai contoh, *tweet* dari pengguna "dinsskuyyy" memiliki nilai sentimen sebesar 0.25, sehingga diberi label Positif, sedangkan *tweet* dari pengguna "Rafitheda" memiliki nilai sentimen sebesar -0.4, sehingga diberi label Negatif. Untuk memberikan pemahaman yang lebih jelas mengenai distribusi hasil pelabelan, perbandingan jumlah data pada setiap kategori ditampilkan dalam sebuah diagram. Diagram tersebut menunjukkan proporsi data untuk masing-masing kategori, seperti yang terlihat pada Gambar 2.



Gambar 2. Hasil Pelabelan

Gambar 2 menunjukan hasil distribusi jumlah *tweet* berdasarkan sentimen positif dan negatif terkait topik *Overclaim Skincare* yang divisualisasikan menggunakan diagram batang. Hasilnya menunjukkan Data yang berhasil dikumpulkan sebelumya berjumlah 7.774 *tweet*. Setelah melalui tahap penghapusan duplikasi, jumlah data yang valid mencapai 5.559 *tweet*. Dari jumlah tersebut, distribusi sentimen menunjukkan dominasi sentimen negatif terhadap topik *Overclaim Skincare*. Hal ini terlihat dari 4.281 *tweet* yang bernada negatif, sementara *tweet* dengan sentimen positif hanya berjumlah 1.275 *tweet*. Visualisasi ini mengindikasikan bahwasebagian besar opini yang dikumpulkan memiliki kecenderungan sentimen negatif terhadap *Overclaim Skincare*, dengan perbedaan jumlah yang signifikan antara kedua kategori yang menegaskan dominasi sentimen negatif dalam dataset ini.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

3.3 Preprocessing

Setelah mendapatkan data yang valid, langkah selanjutnya adalah tahap *preprocessing* atau pemrosesan teks. Tahap ini sangat penting dalam mengelola dan menganalisis data berbasis teks, terutama untuk mendukung analisis sentimen terkait topik *overclaim skincare*. *Preprocessing* berperan penting dalam *Natural Language Processing* (NLP), yang memungkinkan komputer untuk memahami, menganalisis, dan menghasilkan bahasa manusia secara bermakna. Proses *preprocessing* melibatkan beberapa langkah, di antaranya, seperti *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *filtering*, dan *stemming*, Teknik-teknik ini memungkinkan komputer mengolah teks secara lebih efektif untuk berbagai tujuan analisis. Hasil dari proses *preprocessing* sebagai berikut:

3.3.1 Cleansing

Cleansing dalam analisis sentimen adalah proses merapikan teks dengan menghapus elemen yang tidak relevan, seperti tanda baca, karakter spesial, angka, URL, mention, hashtag, kata umum seperti "di" atau "ke," serta spasi berlebih. Hasil *cleansing* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Proses *Cleansing*.

Tweet	*Cleansing*
Guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim https://t.co/MlVbRoZBvM	Guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim
bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang https://t.co/JL2yraE3ub	bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang

Tabel 3 menunjukkan hasil *cleansing* teks *tweet*. Kolom pertama berisi teks asli dengan elemen tidak relevan seperti URL, sementara kolom kedua menampilkan teks yang telah dibersihkan. Proses ini menjadikan teks lebih rapi dan siap untuk analisis sentimen

3.3.2 Case Folding

Case folding adalah teknik dalam pemrosesan teks yang mengubah semua huruf dalam teks menjadi huruf kecil. Hasil proses *casefolding* disajikan pada Tabel 4.

Tabel 4. Hasil Proses *Case Folding*.

Cleansing	*Case Folding*
Guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim	guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim
bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang	bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang

Tabel 4 menunjukkan hasil dari proses *case folding* pada teks yang telah dibersihkan. Kolom pertama menampilkan teks setelah proses *cleansing*, sementara kolom kedua menunjukkan teks yang telah diubah seluruhnya menjadi huruf kecil. Teknik ini bertujuan untuk menghilangkan perbedaan kapitalisasi huruf, sehingga mempermudah analisis dan membuat pencarian lebih konsisten.

3.3.3 Tokenizing

Pada tahap ini, digunakan *library NLTK* untuk melakukan tokenisasi, yang bertujuan mengubah teks menjadi unit-unit terkecil yang disebut token. Hasil *tokenizing* dapat dilihat pada Tabel 5.

Tabel 5. Hasil Proses *Tokenizing*.

Case Folding	*Tokenizing*
guysss please be aware yaa karena zaman sekarang tuh udah banyak bangett skincare skincare yang pada overclaim	['guysss', 'please', 'be', 'aware', 'yaa', 'karena', 'zaman', 'sekarang', 'tuh', 'udah', 'banyak', 'bangett', 'skincare', 'skincare', 'yang', 'pada', 'overclaim']
bahayaaa nih makin hari makin banyak skincare overclaim yang terkuak jangan sampe ketipu guys harus pinter milih skincare sekarang	['bahayaaa', 'nih', 'makin', 'hari', 'makin', 'banyak', 'skincare', 'overclaim', 'yang', 'terkuak', 'jangan', 'sampe', 'ketipu', 'guys', 'harus', 'pinter', 'milih', 'skincare', 'sekarang']

Tabel 5 menunjukkan hasil dari proses *tokenizing* pada teks yang sudah melalui tahap *case folding*. Kolom pertama berisi teks yang telah diubah menjadi huruf kecil, sedangkan kolom kedua menampilkan teks yang telah dipisahkan menjadi unit-unit kecil atau token, seperti kata atau frasa.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

3.3.4 Stopword Removal

Stopwords dalam analisis sentimen adalah kata-kata umum yang tidak memberikan kontribusi signifikan terhadap makna atau emosi teks. Hasil proses dari *stopword* dapat dilihat pada Tabel 6.

Tabel 6. Hasil Proses *Stopword Removal*.

Tokenizing	***Stopword Removal***
['guysss', 'please', 'be', 'aware', 'yaa', 'karena', 'zaman', 'sekarang', 'tuh', 'udah', 'banyak', 'bangett', 'skincare', 'skincare', 'yang', 'pada', 'overclaim']	['guysss', 'please', 'aware', 'yaa', 'zaman', 'tuh', 'udah', 'bangett', 'skincare', 'skincare', 'yang', 'overclaim']
['bahayaaa', 'nih', 'makin', 'hari', 'makin', 'banyak', 'skincare', 'overclaim', 'yang', 'terkuak', 'jangan', 'sampe', 'ketipu', 'guys', 'harus', 'pinter', 'milih', 'skincare', 'sekarang']	['bahayaaa', 'nih', 'skincare', 'overclaim', 'yang', 'terkuak', 'sampe', 'ketipu', 'guys', 'pinter', 'milih', 'skincare']

Tabel 6 menunjukkan hasil *stopword removal* setelah *tokenisasi*, di mana kata-kata umum yang tidak relevan seperti "be", "please", dan "karena" dihapus, sehingga hanya token yang lebih bermakna yang tersisa untuk analisis.sehingga algoritma dapat fokus pada kata-kata yang lebih relevan

3.3.5 Stemming

Stemming adalah metode dalam pemrosesan teks yang digunakan untuk mengubah kata-kata menjadi bentuk dasarnya atau "akar" kata. Hasil proses dari *Stemming* dapat dilihat pada Tabel 7.

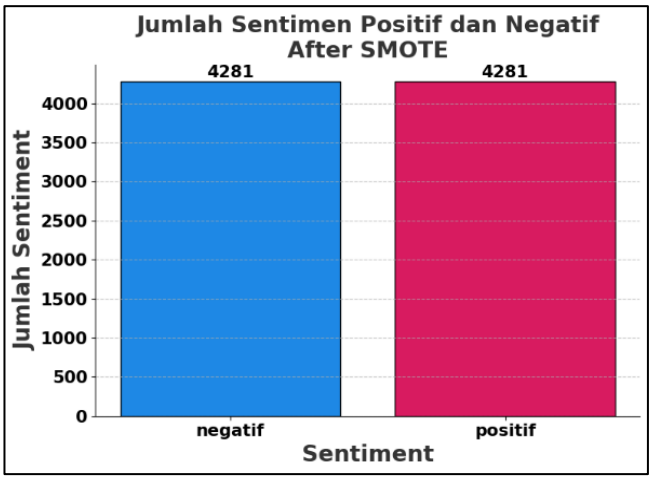
Tabel 7. Hasil Proses *Stemming*.

Stopword Removal	***Stemming***
['guysss', 'please', 'aware', 'yaa', 'zaman', 'tuh', 'udah', 'bangett', 'skincare', 'skincare', 'yang', 'overclaim']	guysss please aware yaa zaman tuh udah bangett skincare skincare yang overclaim
['bahayaaa', 'nih', 'skincare', 'overclaim', 'yang', 'terkuak', 'sampe', 'ketipu', 'guys', 'pinter', 'milih', 'skincare']	bahayaaa nih skincare overclaim yang kuak sampe tipu guys pinter milih skincare

Tabel 7 menunjukkan hasil dari proses *stemming* setelah penghapusan *stopwords*. Pada kolom pertama, terlihat token yang telah dihapus *stopwords*-nya. Pada kolom kedua, kata-kata tersebut telah diproses melalui *stemming*, yang mengubah kata-kata seperti "kuak" menjadi bentuk dasar "kuak" dan "terkuak" menjadi "terkuak".

3.4 SMOTE

Setelah tahap *preprocessing*, langkah berikutnya adalah mengatasi ketidakseimbangan data sentimen.Data awal menunjukkan dominasi sentimen negatif dengan 4.281 *tweet* dibandingkan sentimen positif yang hanya 1.275 *tweet*. Ketidakseimbangan ini dapat menyebabkan bias model terhadap kelas *mayoritas*.Untuk mengatasinya, diterapkan teknik SMOTE *(Synthetic Minority Oversampling Technique)*, yang mensintesis sampel baru untuk kelas *minoritas* berdasarkan tetangga terdekat. Hasil optimasi SMOTE dilihat pada Gambar 3.



Gambar 3. Optimasi SMOTE.

Berdasarkan Gambar 3 yang menunjukan hasil optimasi SMOTE, yang dimana teknik ini menyeimbangkan distribusi data menjadi 4.281 *tweet* untuk masing-masing kelas, sehingga model dapat mempelajari pola dari kedua kelas secara lebih akurat., yang menunjukkan distribusi data yang lebih seimbang, meningkatkan akurasi prediksi tanpa bias pada kelas *mayoritas*.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

3.5 Klasifikasi Algoritma SVM, Random Forest, KNN

Langkah berikutnya adalah tahap klasifikasi, yang bertujuan untuk mengevaluasi kinerja model sebelum dan sesudah diterapkannya teknik *Synthetic Minority Oversampling Technique* (SMOTE) sebagai upaya mengatasi ketidakseimbangan kelas pada dataset. Tiga algoritma klasifikasi yang diuji adalah *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest Neighbors* (KNN). Kinerja model dinilai menggunakan metrik evaluasi seperti *Accuracy, Precision, Recall, dan F1-Score*. Dataset yang digunakan memiliki distribusi kelas yang tidak seimbang, di mana kelas *minoritas* memiliki jumlah data yang jauh lebih kecil dibandingkan dengan kelas *mayoritas.* Hasil klasifikasi sebelum dan sesudah penerapan SMOTE dapat dilihat pada Tabel 8 dan Tabel 9.

Tabel 8. Hasil Klasifikasi *Before* SMOTE

Model	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	Negatif	91%	90%	99%	94%
	Positif		95%	65%	77%
Random Forest	Negatif	95%	94%	99%	97%
	Positif		98%	79%	87%
KNN	Negatif	80%	81%	96%	88%
	Positif		71%	29%	41%

Berdasarkan hasil klasifikasi pada Tabel 8, Sebelum penerapan SMOTE, hasil klasifikasi menunjukkan kinerja yang berbeda pada masing-masing model. Model SVM memiliki akurasi yang baik pada kelas negatif 91% dengan *recall* yang sangat tinggi 99%, namun *precision* dan *F1-score* pada kelas positif masih rendah, yaitu 65% dan 70%. Hal ini menunjukkan bahwa meskipun model berhasil mendeteksi sebagian besar kelas negatif dengan baik, prediksi untuk kelas positif masih kurang akurat. *Random Forest* juga menunjukkan akurasi yang baik, dengan 95% untuk kelas negatif dan 98% untuk kelas positif, namun *precision* untuk kelas positif masih 79% dan *recall* 87%. Meskipun performanya cukup baik, keduanya masih dapat ditingkatkan lebih lanjut, terutama untuk kelas positif. Sedangkan KNN menunjukkan kinerja yang paling buruk, dengan akurasi hanya 80% untuk kelas negatif dan 71% untuk kelas positif. *Precision* pada kelas positif hanya 29%, *recall* 41%, dan *F1-score* yang sangat rendah 34%, yang menandakan ketidakmampuan model ini dalam mendeteksi kelas positif dengan baik.

Tabel 9. Hasil Klasifikasi *After* SMOTE

Model	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	Negatif	97%	96%	98%	97%
	Positif		97%	96%	97%
Random Forest	Negatif	98%	97%	99%	98%
	Positif		99%	97%	98%
KNN	Negatif	84%	86%	81%	84%
	Positif		82%	86%	84%

Setelah penerapan SMOTE, kinerja model mengalami peningkatan signifikan. SVM kini memiliki akurasi yang lebih tinggi pada kedua kelas, mencapai 97% pada kelas negatif dan 97% pada kelas positif. *Precision* 96%, *recall* 97%, dan *F1-score* 97% untuk kelas positif juga mengalami peningkatan yang signifikan, menunjukkan bahwa SMOTE berhasil meningkatkan deteksi kelas positif. *Random Forest* juga menunjukkan hasil yang luar biasa setelah SMOTE, dengan akurasi 98% pada kelas negatif dan 99% pada kelas positif. *Precision* 97% dan *recall* 98% pada kelas positif pun semakin seimbang, yang menunjukkan peningkatan kinerja yang sangat baik. Sementara itu, KNN mengalami peningkatan di kelas negatif dengan akurasi 84% dan *precision* 86%, meskipun akurasi dan F1-score pada kelas positif masih lebih rendah dibandingkan model lainnya, dengan akurasi 82% dan *F1-score* 85%. Secara keseluruhan, penerapan SMOTE terbukti efektif dalam meningkatkan keseimbangan dan kinerja model, terutama pada kelas positif, yang sebelumnya memiliki performa yang kurang optimal.

3.6 Evaluasi

Setelah proses klasifikasi menggunakan model SVM, *Random Forest*, dan KNN, selanjutnya model dievaluasi menggunakan *confusion matrix*. Evaluasi ini memberikan gambaran yang lebih mendalam mengenai kinerja model dalam hal prediksi kelas, termasuk jumlah prediksi yang benar *true positive* dan *true negative* serta jumlah prediksi yang salah *false fositive* dan *false negative*. Dengan menggunakan *confusion matrix*, kita dapat menganalisis lebih lanjut kemampuan model dalam mendeteksi kedua kelas, serta mengukur *metrik-metrik* penting lainnya seperti *precision, recall*, dan *F1-score*. *Confusion Matrix Before* dan *After* SMOTE dapat dilihat pada tabel 9 dan 10.

Tabel 10. *Confusion Matrix Before* SMOTE

Model	*Actual Class*	*Prediction Class*	
		Prediction Negatif	*Prediction Positif*
SVM	Negatif	840	9

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

Random Forest	Positif	92	171
	Negatif	844	5
	Positif	55	208
KNN	Negatif	818	31
	Positif	188	75

Berdasarkan Tabel 10 model *support vector machine* memprediksi 840 data negatif benar dan 171 data positif benar, dengan 92 *False Negatives* dan 9 *False Positives*. Model *Random Forest* memprediksi 844 data negatif benar dan 208 data positif benar, dengan 55 *False Negatives* dan 5 *False Positives*. Model *K-Nearest Neighbors* memprediksi 818 data negatif benar dan 75 data positif benar, dengan 188 *False Negatives* dan 31 *False Positives*. Secara keseluruhan, sebelum SMOTE, ketiga model cenderung lebih baik dalam mendeteksi kelas negatif, namun memiliki kelemahan signifikan dalam mengenali kelas positif

Tabel 10. *Confusion Matrix After* SMOTE

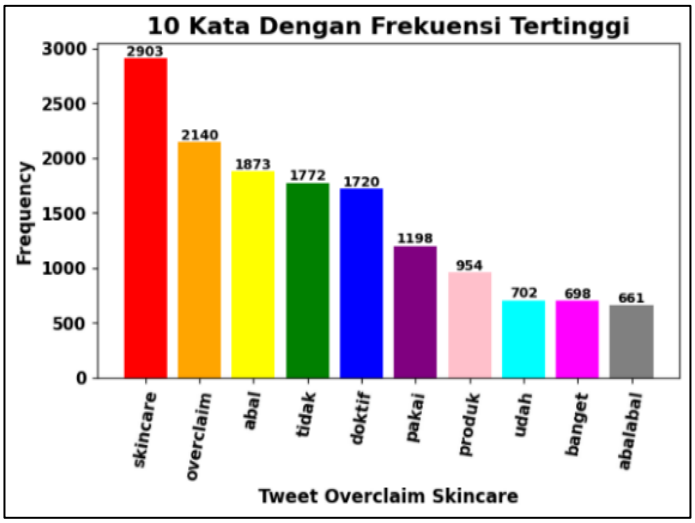
Model	***Actual Class***	***Prediction Class***	
		Prediction Negatif	***Prediction Positif***
SVM	Negatif	841	21
	Positif	33	818
Random Forest	Negatif	1276	7
	Positif	44	1242
KNN	Negatif	702	160
	Positif	115	736

Setelah penerapan SMOTE, performa model meningkat signifikan, terutama dalam mendeteksi kelas positif. Model SVM menunjukkan kinerja yang baik dengan 841 prediksi negatif dan 818 prediksi positif yang benar, serta kesalahan yang lebih kecil 21 *False Positives* dan 33 *False Negatives*. Model *Random Forest* menunjukkan performa terbaik dengan 1,276 prediksi negatif dan 1,242 prediksi positif yang benar, hanya dengan 7 *False Positives* dan 44 *False Negatives*. Model KNN memiliki akurasi yang cukup baik dengan 702 prediksi negatif dan 736 prediksi positif yang benar, meskipun masih memiliki 160 *False Positives* dan 115 *False Negatives*. Secara keseluruhan, penerapan SMOTE *Synthetic Minority Over-sampling Technique* terbukti meningkatkan kinerja model dengan memperbaiki distribusi prediksi, terutama pada kelas *minoritas* dan mengalami peningkatan signifikan dalam *metrik* evaluasi seperti *precision, recall*, dan *F1-score*.

3.7 Visualisasi

3.7.1 Frekuensi Kata

Frekuensi Kata menunjukkan 10 kata dengan frekuensi tertinggi dalam *tweet* tentang "*Overclaim Skincare*.". Dapat dilihat pada Gambar 6.



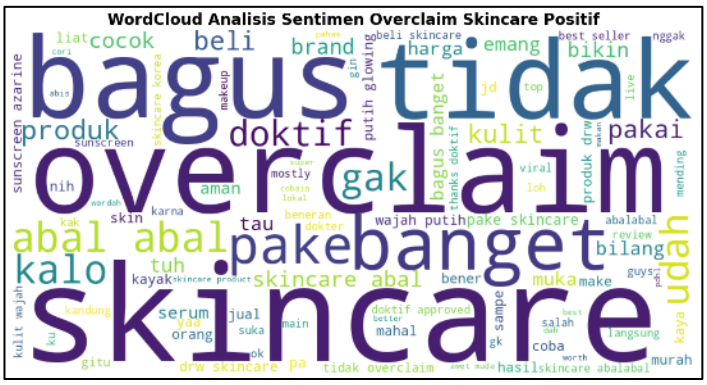
Gambar 6. Hasil *Frekuensi* Kata

Berdasarkan Gambar 6 yang menjukan *Frekuensi* Kata, Grafik menunjukkan kata-kata paling sering muncul dalam tweet tentang overclaim skincare, dengan "skincare" (2903) dan "overclaim" (2140) mendominasi. Kata seperti "abal", "dokter", dan "produk" mencerminkan kekhawatiran akan kualitas dan keaslian produk, menggambarkan tema utama percakapan.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

3.7.2 Visualiasi WordCloud

WordCloud digunakan untuk menampilkan kata-kata yang paling sering muncul dalam data teks. Kata dengan frekuensi tinggi ditampilkan lebih besar, membantu mengidentifikasi tema utama dalam *tweet*, baik untuk sentimen positif maupun negatif. Hasil *wordcloud* sapat dilihat pada Gambar 4 dan Gambar 5.



Gambar 4. *WordCloud* Positif

Berdasarkan Gambar 4, word cloud untuk sentimen negatif menunjukkan kata-kata seperti "*overclaim*," "abal," "doktif," "skincare," dan "produk," yang mencerminkan kritik terhadap klaim yang tidak sesuai atau tidak aman. Kata-kata seperti "dokter," "uji lab," dan "BPOM" menunjukkan keraguan terhadap keraguan klaim.

Gambar 5. *WordCloud* Negatif

Berdasarkan Gambar 5 yang menjukan wordcloud positif, kata-kata seperti "bagus," "aman," "cocok," "*skincare* lokal," dan "*glowing*" menunjukkan penilaian baik terhadap produk yang berkualitas dan sesuai klaim. Analisis ini mencerminkan persepsi konsumen tentang klaim produk *skincare*, dengan fokus pada kritik terhadap *overclaim* dan apresiasi untuk produk yang memenuhi standar.

4. KESIMPULAN

Berdasarkan hasil penelitian, penelitian ini memberikan pemahaman mengenai perbandingan kinerja algoritma *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest Neighbors* (KNN) dalam analisis sentimen terhadap klaim berlebihan *overclaim* produk skincare pada media sosial X. Hasil penelitian menunjukkan bahwa sebelum penerapan SMOTE, *Random Forest* menunjukkan kinerja terbaik dengan akurasi 95% pada kelas negatif dan performa baik pada kelas positif, diikuti oleh SVM dengan akurasi 91% pada kelas negatif meskipun kurang optimal pada sentimen positif. KNN memiliki akurasi terendah, yaitu 80%. Setelah penerapan SMOTE, kinerja model meningkat, dengan SVM mencapai akurasi 97% dengan *precision* dan *recall* yang seimbang, *Random Forest* tetap menunjukkan performa terbaik dengan akurasi 98%, dan KNN mengalami peningkatan menjadi 84%. Dalam penelitian ini, *Random Forest* terbukti sebagai algoritma terbaik, menunjukkan performa tertinggi baik sebelum maupun setelah penerapan SMOTE. Penelitian ini juga menunjukkan bahwa analisis sentimen berbasis *machine learning* dapat membantu industri *skincare* mendeteksi klaim berlebihan yang berpotensi merugikan reputasi produk. Oleh karena itu, perusahaan perlu berhati-hati dalam membuat klaim yang tidak dapat dibuktikan. Penelitian ini memiliki beberapa keterbatasan yang dapat diperbaiki. Pertama, penggunaan dataset terbatas yang berjumlah 7.774 *tweet* yang dikumpulkan antara Oktober hingga November 2024 dapat diperluas dengan dataset yang lebih besar dan lebih beragam untuk meningkatkan *generalisasi* model. Kedua, meskipun SMOTE digunakan untuk menangani ketidakseimbangan data, teknik lain seperti ADASYN *(Adaptive Synthetic Sampling)* atau pengurangan kelas negatif dapat memberikan hasil yang lebih optimal.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

REFERENCES

[1] P. Rachmawati, "Edukasi Terkait Keamanan Kosmetik Kepada Masyarakat," *MitraMas: Jurnal Pengabdian dan Pemberdayaan Masyarakat*, vol. 1, no. 2, pp. 101–113, 2023, doi: 10.25170/mitramas.v1i2.4308.

[2] H. Hartanto and C. Wilda Meutia Syafiina, "Efektivitas Perlindungan Konsumen Terhadap Produk Kosmetik Yang Tidak Memiliki Izin Edar Balai Besar Pengawas Obat Dan Makanan Diy (Dalam Perspektif Hukum Pidana)," *Jurnal Meta-Yuridis*, vol. 4, no. 1, pp. 54–72, 2021, doi: 10.26877/m-y.v4i1.6765.

[3] A. A. P. Kuncoro and M. Syamsudin, "Perlindungan Konsumen terhadap Overclaim Produk Skincare," *Prosiding Seminar Hukum Aktua*, vol. Vol. 2 No., no. September, p. 82, 2024.

[4] S. Lutfiani, R. Astuti, and Fadhil M Basysyar,M,Kom, "Analisis Sentimen Pengaruh Media Sosial Terhadap Minat Beli Skincare Pada Remaja Di Indonesia Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 2957–2961, 2024, doi: 10.36040/jati.v8i3.9614.

[5] A. N. Putri and R. Apriani, "Perlindungan Konsumen Atas Predaran Skincare Yang Belum Mendapat Izin Edar Dari Bpom," *Jurnal Justitia: Jurnal Ilmu Hukum dan Humaniora*, vol. 9, no. 3, pp. 1227–1233, 2022, doi.org/10.31604/justitia.v9i3.1227-1233.

[6] T. Intensi, M. Ulang, and P, "Pengaruh Overclaim Produk, Kesadaran Merek, Kepuasan Konsumen, Loyalitas Konsumen, Terhadap Intensi Membeli Ulang Produk Skincare Skintific Padamahasiswa Aktif Universitas Riau Kepulauan," *Bening Journal*, vol. 11, no. 1, 2024, doi.org/10.33373/bening.v11i1.6319.

[7] M. Manik, P. A. Sipahutar, and M. R. A. Putra, "Tangung Jawab Pelaku Usaha atas Overclaim Produk Skincare di Media Sosial," *Madani: Jurnal Ilmiah Multidisiplin*, vol. 2, no. 10, pp. 663–668, 2024, doi.org/10.5281/zenodo.14185081.

[8] Nabilla Dhinggar Arumbi, Sapto Hermawan, and Asianto Nugroho, "Tanggung Jawab Pelaku Usaha Atas Overclaim Sun Protection Factor (SPF) Pada Produk Tabir Surya X," *Amandemen: Jurnal Ilmu pertahanan, Politik dan Hukum Indonesia*, vol. 1, no. 2, pp. 25–34, 2024, doi: 10.62383/amandemen.v1i2.127.

[9] A. A. L. Widawati and M. Elbana, "Kajian Litelatur Review Krisis Komunikasi Hotto Purto pada Kasus Overclaim dalam Menjaga Citra Perusahaan," *Jurnal Penelitian Inovatif*, vol. 4, no. 1, pp. 113–120, 2024, doi: 10.54082/jupin.262.

[10] H. Harnelia, "Analisis Sentimen Review Skincare Skintific Dengan Algoritma Support Vector Machine (Svm)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4095.

[11] A. F. Setyaningsih, D. Septiyani, and S. R. Widiasari, "Implementasi Algoritma Naïve Bayes untuk Analisis Sentimen Masyarakat pada Twitter mengenai Kepopuleran Produk Skincare di Indonesia," *Jurnal Teknologi Informatika dan Komputer*, vol. 9, no. 1, pp. 224–235, 2023, doi: 10.37012/jtik.v9i1.1409.

[12] M. H. Wicaksono, M. D. Purbolaksono, and S. Al Faraby, "Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily," *eProceedings of Engineering*, vol. 10, no. 3, pp. 3591–3600, 2023.

[13] N. Resti Wardani, S. Saepudin, and C. Warman, "Sentimen Analisis Kegiatan Trading Pada Ap-likasi Twitter dengan Algoritma SVM, KNN Dan Random Forrest," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 6, no. 2, pp. 863–870, 2022, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

[14] M. N. Muttaqin and I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," *UNNES Journal of Mathematics*, vol. 10, no. 2, pp. 22–27, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm

[15] R. R. S. Putri Kumala Sari, "Komparasi Algoritma Support Vector Machine Dan Random Forest Untuk Analisis Sentimen Metaverse," *Jurnal MNEMONIC*, vol. 7, no. 1, pp. 31–39, 2024, doi.org/10.36040/mnemonic.v7i2.

[16] C. P. Yanti, N. W. Eva Agustini, N. L. W. Sri Rahayu Ginantra, and D. A. Putri Wulandari, "Perbandingan Metode K-NN Dan Metode Random Forest Untuk Analisis Sentimen pada Tweet Isu Minyak Goreng di Indonesia," *Jurnal Media Informatika Budidarma*, vol. 7, no. 2, p. 756, 2023, doi: 10.30865/mib.v7i2.5900.

[17] I. Septiana and D. Alita, "Perbandingan Random Forest dan SVM dalam Analisis Sentimen Quick Count Pemilu 2024," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 3, pp. 224–233, 2024, doi: 10.30591/jpit.v9i3.6640.

[18] H. Ali, N. Hendrastuty, "Comparison Of Naïve Bayes Classifier, Support Vector Machine, Random Forest Algorithms For Public Sentiment Analysis Of Kip-K Program On Twitter," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 6, pp. 1701–1712, 2024, doi.org/10.52436/1.jutif.2024.5.6.4030.

[19] P. Cahyani and L. Abdillah, "Perbandingan Performa Algoritma Naïve Bayes , SVM dan Random Forest : Studi Kasus Analisis Sentimen Pengguna Sosial Media X," *Jurnal Sains dan Teknologi*, vol. 11, no. 02, pp. 12–21, 2024, doi.org/10.53008/kalbiscientia.v11i02.3624.

[20] D. A. Fitri and Damayanti, "Komparasi algoritma random forest classifier dan support vector machine untuk sentimen masyarakat terhadap pinjaman online di media sosial," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 2018–2029, 2024, doi.org/10.29100/jipi.v9i4.5608.

[21] A. Baita, Y. Pristyanto, and N. Cahyono, "Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (Svm) Dan K-Nearest Neighbor (Knn)," *Information System Journal (INFOS)*, vol. 4, no. 2, pp. 42–42, 2021, doi.org/10.24076/infosjournal.2021v4i2.687.

[22] D. S. Ningsih and R. R. Suryono, "Comparison of Naïve Bayes and Information Gain Algorithms in Cyberbullying Sentiment Analysis on Twitter Perbandingan Algoritma Naïve Bayes Dan Information Gain," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 4, pp. 1085–1091, 2024, doi.org/10.52436/1.jutif.2024.5.4.1908.

[23] O. I. Gifari, Muh. Adha, F. Freddy, and F. F. S. Durrand, "Film Review Sentiment Analysis Using TF-IDF and Support Vector Machine," *Journal of Information Technology*, vol. 2, no. 1, pp. 36–40, 2022, doi.org/10.46229/jifotech.v2i1.330.

[24] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, 2021, doi: 10.29207/resti.v5i4.3308.

[25] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, 2023, doi: 10.33365/jatika.v4i1.2468.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2390−2402
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6782

[26] J. Anggraini and D. Alita, "Implementasi Metode SVM Pada Sentimen Analisis Terhadap Pemilihan Presiden (Pilpres) 2024 Di Twitter," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 2, pp. 102–111, 2024, doi: 10.30591/jpit.v9i2.6560.
[27] Yulistiani and Styawati, "Analisis Sentimen Terhadap Calon Presiden Indonesia 2024 dengan Metode Extreme Gradient Boosting (XGBOOST)," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 3, pp. 322–328, 2024, doi: 10.30591/jpit.v9i3.6127.
[28] N. Hendrastuty, A. Rahman Isnain, and A. Yanti Rahmadhani, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *Jurnal Informatika: Jurnal pengembangan IT*, vol. 6, no. 3, pp. 150–155, 2021.
[29] "Eskiyaturrofikoh" and R. R. 'Suryono, "Analisis Sentimen Aplikasi X Pada Google Play Store Menggunakan Algoritma Naïve Bayes Dan Support Vector Machine (Svm)," *JIPI(Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 3, pp. 1408–1419, 2024, doi.org/10.29100/jipi.v9i3.
[30] D. Kurniawan, M. Najib, and D. Satria, "Analisis Sentimen Opini Publik Tentang Gempa Megathrust di Indonesia Menggunakan Metode Support Vector Machine dan Naïve Bayes," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, 2024, doi: 10.47065/bits.v6i3.6213.
[31] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 13, no. 1, pp. 84–93, 2023, doi: 10.34010/jamika.v13i1.9424.
[32] Tommy Suhendra, B. Intan, and A. T. Martadinata, "Analisis Sentimen Pengguna Aplikasi Netflix Pada Ulasan Google Playstore Menggunakan Metode Naïve Bayes," *ESCAF 3rd*, vol. 2, no. 2, pp. 1011–1022, 2024, doi: 10.47065/bits.v6i2.5528.
[33] T. Tinaliah and T. Elizabeth, "Analisis Sentimen Ulasan Aplikasi PrimaKu Menggunakan Metode Support Vector Machine," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 4, pp. 3436–3442, 2022, doi: 10.35957/jatisi.v9i4.3586.
[34] R. S. Arischo and D. Damayanti, "Analisis Sentimen Pinjaman Online di Twitter dengan Metode Naive Bayes Classifier dan SVM," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, p. 1120, 2024, doi: 10.30865/mib.v8i2.7406.
[35] M. Azhari and Parjito, "Analisis Sentimen Opini Publik Program Makan Siang Gratis dengan Random Forest Pada Media X," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1932–1942, 2024, doi: 10.47065/bits.v6i3.6423.
[36] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," *Jurnal Informatika Upgris*, vol. 7, no. 1, pp. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099.
[37] A. D. Adhi Putra, "Sentiment Analysis on User Reviews of the Bibit and Bareksa Application with the KNN Algorithm," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi:10.35957/Jatisi.V8i2.962.
[38] R. Aryanti, T. Misriati, and A. Sagiyanto, "Analisis Sentimen Aplikasi Primaku Menggunakan Algoritma Random Forest dan SMOTE untuk Mengatasi Ketidakseimbangan Data," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 1, pp. 218–227, 2023, doi: 10.47065/josyc.v5i1.4562.
[39] R. Nurhidayat and N. Hendrastuty, "Analisis Sentimen Komentar Media Sosial Twitter Terhadap Tes CPNS dengan Algoritma Naive Bayes," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1477–1489, 2024, doi: 10.47065/bits.v6i3.6148.