Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) pada Analisis Sentimen Capcut

Charles Zai[1], Auliya Rahman Isnain[2]
[1,2]Universitas Teknokrat Indonesia,
Jl. ZA. Pagar Alam No.9-11, Labuan Ratu, Kec. Kedaton, Kota Bandar Lampung, Lampung 35132
E-mail: charles_zai@teknokrat.ac.id[1], auliyarahman@teknokrat.ac.id[2]

Abstract – Sentiment analysis on capcut app reviews is essential in understanding user opinions. Reviews provided by users can provide valuable insights for developers to improve the quality and performance of the application. Data collection is taken from Google PlayStore as much as 3000 data by performing srcapping techniques, then the data is preprocessed so that the data used in this study is 2993 data, and then comparing classification models using Naïve Bayes and Support vector Machine (SVM). These two algorithms are classification methods that are often used to process text data with a good level of accuracy. The naïve bayes algorithm has 78% accuracy and the SVM algorithm has 81% accuracy, but the recall and f1-socre results are low. This indicates that the machine still dominates in the majority label because the amount of negative sentiment is 1867 data, while positive sentiment gets 1126 data. Therefore, researchers apply SMOTE to overcome the imbalance between classes, with the ability to increase the representation of minority classes. The results of SMOTE optimization on the SVM algorithm are superior with an accuracy value of 86%, precision 85%, recall 85%, f1-score 85%, while the Naïve Bayes algorithm has an accuracy value of 81%, precision 84%, recall 79%, f1-score 81%. The positive sentiment wordcloud visualization results refer to user satisfaction while the negative sentiment is user inconvenience in performing the video editing process.

Keywords - sentiment analysis, Naive Bayes, SVM, Capcut, SMOTE.

Intisari – Analisis sentimen pada ulasan aplikasi capcut sangat penting dalam memahami opini pengguna. Ulasan yang diberikan oleh pengguna dapat memberikan wawasan berharga bagi pengembang untuk meningkatkan kualitas dan kinerja aplikasi. Pengumpulan data diambil dari Google PlayStore sebanyak 3000 data dengan melakukan teknik srcapping, selanjutnya data di preprocessing sehingga data yang digunakan pada penelitian ini 2993 data, dan selanjutnya melakukan komparasi model klasifikasi dengan menggunakan Naïve Bayes dan Support vector Machine (SVM). Kedua algoritma ini merupakan metode klasifikasi yang sering digunakan untuk mengolah data berupa teks dengan tingkat akurasi yang baik. Hasil komparasi algoritma naïve bayes memiliki akurasi 78% dan algoritma SVM memperoleh akurasi 81%, namun hasil recall dan f1-socre rendah. Hal ini menandakan bahwa mesin masih mendominasi pada label mayoritas dikarenakan jumlah pada sentimen negatif 1867 data, sedangkan sentimen positif mendapatkan 1126 data. Oleh karena itu peneliti menerapkan SMOTE untuk mengatasi ketidakseimbangan antar kelas, dengan kemampuan meningkatkan representasi kelas minoritas. Hasil optimasi SMOTE pada algoritma SVM lebih unggul dengan nilai akurasi 86%, precision 85% , recall 85%, f1-score 85%, sedangkan algoritma Naïve Bayes nilai akurasi 81%, precision 84%, recall 79%, f1-score 81%. Hasil visualisasi wordcloud sentimen positif mengacu pada kepuasan pengguna sedangkan sentimen negatif ketidaknyamanan pengguna dalam melakukan proses pengeditan video.

Kata Kunci – analisis sentimen, Naive Bayes, SVM, Capcut, SMOTE

I. PENDAHULUAN

Di kehidupan sehari-hari, teknologi internet yang telah menjadi salah satu kebutuhan yang tak terpisahkan dari aktivitas manusia[1], dengan keberadaan internet telah mengubah cara kita berkomunikasi, bekerja, dan mengakses informasi [2]. Di indonesia, telah mengalami pertumbuhan internet yang sangat pesat dalam beberapa tahun terakhir. Berdasarkan survei APJII (Asosiasi Penyelenggara Jasa Internet Indoneisa), jumlah penggunaan internet di indoensia meningkat dari 210,03 juta pada tahun 2021--2022 menjadi 221,56 juta pada tahun 2023-2024 [3]. Pertumbuhan ini mencerminkan peningkatan aksesibilitas dan adopsi teknologi internet di lapisan seluruh masyarakat.

Peningkatan jumlah pengguna internet ini memiliki dampak signifikan terhadap penggunaan aplikasi digital[4]. Semakin banyaknya pengguna internet, semakin banyak orang yang dapat mengakses dan memanfaatkan berbagai aplikasi untuk kebutuhan sehari-hari, baik itu untuk komunikasi, hiburan, produktivitas, maupun kreativitas [5]. Salah satu aplikasi yang menonjol dalam kategori kreatif adalah capcut, sebuah aplikasi aplikasi pengeditan video yang dimiliki oleh perusahaan tiongkok yaitu ByteDance Pte.Ltd, dimana aplikasi tersebut tersedia untuk di unduh secara gratis di Google Playstore [6]. Aplikasi ini menyediakan untuk membuat video berkualitas tinggi dengan mudah dan cepat. Capcut telah mendapatkan popularitas yang luar biasa, terutama di kalangan kreator konten yang aktif di platform media sosial seperti TikTok, Instagram, dan YouTube. Aplikasi ini menawarkan beragam fitur, termasuk pemotongan video, penambahan efek visual, transisi, teks, dan musik, yang semuanya dirancang untuk mendukung kreativitas pengguna dalam menghasilkan konten yang menarik dan profesional [7]. Namun, popularitas capcut juga menjadi tantangan, setiap perubahan dalam aplikasi dapat mempengaruhi pengalaman pengguna dengan sudut pandang yang berbeda, sehingga perlu dilakukan analisis sentimen terhadap ulasan aplikasi capcut.

Tujuan dari penelitian ini untuk menganalisis sentimen terhadap pengguna aplikasi capcut pada ulasan di Google Play Store, untuk mengetahui sentimen pengguna terdahap layanan aplikasi apakah pengguna lebih cenderung berkomentar sentimen positif atau negatif. Selain itu, penelitian ini juga memiliki tujuan untuk melakukan komparasi menggunakan algortima Naive Bayes dan Support Vector Machine (SVM). Selain perbandingan antara kedua algoritma tersebut untuk menentukan algoritma terbaik yang ditunjukkan pada hasil akurasi, precision, recall, f1-score. Penelitian ini juga melakukan metode optimasi SMOTE (Synthetic Minority Oversampling Technique) agar kinerja setiap model klasifikasi yang dilatih menjadi lebih optimal. Penelitian ini juga serta memberikan masukan kepada pengembang aplikasi capcut untuk mengambil langkah yang diperlukan untuk meningkatkan layanan terdahap aplikasi capcut.

Bedasarkan uraian diatas, penelitian ini diharapkan akan menjawab beberapa pertanyaan penting, yang pertama, Bagaimana distribusi sentimen positif dan negatif pada ulasan penggunaan aplikasi capcut. Kedua, Komparasi algoritma Naïve Bayes dan Support Vector machine dengan menentukan algoritma mana kinerjanya lebih baik. Ketiga, Bagaimana pengaruh metode *SMOTE* terhadap kinerja algoritma klasifikasi Naïve Bayes dan Support Vector Machine. Keempat, Apa saran yang dapat diberikan kepada pengembang aplikasi capcut berdasarkan analisis sentimen.

II. SIGNIFIANSI STUDI

A. Penelitian Terdahulu

Beberapa analisis terdahulu untuk dijadikan pedoman atau referensi dalam penulisan pada analisis ini sebagaimana yang dipaparkan pada tabel I.

TABEL I
PENELITIAN TERDAHULU

Nomor	Penulis	Penelitian Terdahulu
1	Moh Khoirul Insan, Umi Hayati dan Odi Nurdiawan [8].	Analisis sentimen ulasan pengguna memakai algoritma Naïve Bayes untuk aplikasi Brimo. Informasi yang diambil dari evaluasi Google Play dari aplikasi Brimo dengan cara pengikisan web yang berat menghasilkan tingkat akurasi 84,52%, tingkat presisi 82,51%, dan tingkat penarikan 87,62%.
2	Friska Aditia Indriyani, Ahmad Fauzi dan Sultan Faisal [3].	Menggunakan Algoritma Naïve Bayes dan SVM, analisis sentimen aplikasi TikTok. Analisis ini memakai informasi yang diperoleh dari evaluasi Google Play Store terhadap aplikasi TikTok dengan kategori yang paling relevan. Dalam penelitian ini, sentimen positif menyumbang 76,7% dari temuan kategorisasi, sementara sentimen negatif menyumbang 23,3%. Hasil akhir metode SVM dengan akurasi 79% lebih tinggi dibanding metode Naïve Bayes dengan akurasi 84%.
3	Antonius Mbay Ndapamuri, Danny Manongga, Ade Iraini [9].	Analisa Ulasan Aplikasi Tripadvisor Dengan Metode SVM, K-Nearest Neighbor, dan Naïve Bayes. Hasil dari evaluasi model SVM dengan akurasi senilai 89.8%, dan model KNN senilai 89.02%, dan Naïve Bayes 88.65%.

B. Kajian Pustaka

Text mining adalah proses pencarian dan evaluasi dataset yang besar berupa teks, dengan tujuan mendapatkan informasi yang bermanfaat di dalam data teks pada topik tertentu [10]. Analisis sentimen adalah bagian dari text mining yang mempelajari teknik untuk menginterpretasikan dan mengevaluasi emosi yang ada pada sebuah teks, tujuan dari analsis sentimen ialah mengelompokkan sentimen pada teks tersebut ke dalam kategori positif, negatif, netral [11]. Pengelompokan sentimen ini dapat digunakan untuk menjalankan klasifikasi pada teks. Dalam proses analisis sentimen, metode klasifikasi yang sering digunakan adalah algoritma Naive Bayes dan algoritma Support Vector Machine (SVM).

Thomas Bayes, seorang ilmuwan inggris, menciptakan metode Naive Bayes yang terkenal untuk kategorisasi analisis sentimen. Pendekatan ini menggunakan perhitungan probabilitas dan statistik. Pendekatan Naive Bayes memakai data historis untuk membuat prediksi tentang masa depan [12]. Untuk mencari perhitungan algoritma Naive Bayes dapat dilihat permasaan 1.

$$P(H|X) = \frac{P(P|H)\,P(H)}{P(X)} \quad (1)$$

Kerterangan:
$P(H|X)$ adalah peluang hipotesis H berdasarkan kondisi X.
X yaitu data laitih dengan kelas (label) yang diketahui.
H yaitu data kelas (label).
$P(H)$ yaitu peluang dari hipotesis dari X yang ditinjau dan.
$P(X)$ yaitu peluang X beradasarkan pada kondisi hipotesis H.

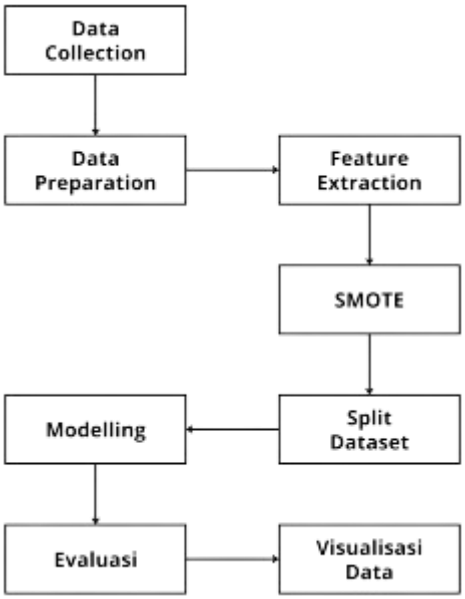
Support Vector Machine (SVM) yaitu algoritma klasifikasi yang bekerja dengan menemukan hyperplane yaitu, batas pemisah yang memiliki margin terbesar antara kelas data yang berbeda. Margin ini mengukur seberapa jauh titik data masing-masing kelas dari hyperplane, Titik-titik yang paling dekat dengan hyperplane dikenal sebagai vektor dukungan. Tujuan utama SVM adalah mencari hyperplane yang memiliki margin maksimum, sehingga

memastikan jarak yang maksimal antara support vector dan hyperplane yang terbentuk Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier [13].

$$f(x) = sign\ (\sum_{i=1}^{n} a_i\ y_i\ K\langle x_i\ x\rangle + b) \quad (2)$$

Keterangan:
$F(X)$ adalah menentukan klasifikasi vektor x ke dalam kelas positif dan negatif.
a_i adalah koefisien lagrange yang dihitung selama proses pelatihan.
y_i adalah kelas target (misalnya -1 untuk negatif dan 1 untuk sentimen positif).
x_i adalah vektor fitur untuk sampel ke – i.
$K\langle x_i\ x\rangle$ adalah kernel yang menghitung jarak antara x_i dan x dalam ruang fitur yang diperluas.
b adalah bias.

C. Metode Penelitan



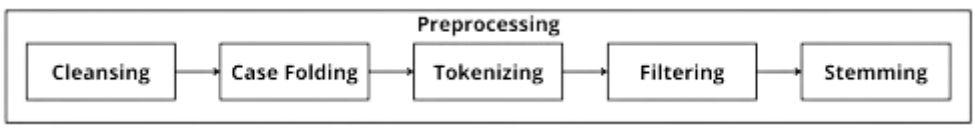
Gambar 1. Alur Penelitian

1. Data Collection (Penumpulan Data)

Data yang dipakai pada analisis ini bersumber dari review aplikasi capcut di google play store. Proses pengambilan dataset melaksanakan teknik web scrapping dengan memanfaatkan bahasa pemrograman python yang ada di platfrom google colab. Data yang telah terkumpul sebanyak 3000 data ulasan yang diambil dari katagori paling *relevant*, lalu data akan disimpan dalam format csv.

2. Data Preparation

Data preparation atau bisa disebut sebagai tahap preprocessing, yang dimana proses ini mempersiapkan data yang bersih dari data yang belum sistematis dan siap dipakai untuk penelitian berikutnya. Adapun Langkah preprocessing yang dilakukan pada analisis ini. Tahapan preprocessing bisa ditinjau pada gambar 2.



Gambar 2. Tahap preprocessing

Setelah melakukan preprocessing data, Lalu diikuti dengan pelabelan data pada ulasan menurut rating atau skor untuk menetapkan sentimen yang di berikan oleh pemakai. Pelabelan ini akan dibagi menjadi 2 katagori, yaitu sentimen positif dan negatif.

3. Feature Extraction

Pada Tahap ini feature extraction memakai TF-IDF. TF-IDF ialah teknik yang dipakai untuk memberi pembobotan kata dalam suatu teks [14]. Pembobotan kata dengan TF-IDF ialah proses yang merubah dokumen teks jadi bentuk numerik berdasarkan bobot dari setiap kata [15]. TF – IDF tersusun dari TF dan IDF. TF yaitu frekuensi dimana sebuah kata muncul pada dokumen, IDF yaitu kebalikan dari frekuensi dokumen [14]. penjumlahan pembobotan TF-IDF bisa ditinjau pada persamaan 3 dan 4.

$$W_{t,d} = TF_{t,d}\ X\ IDF_t \tag{3}$$

$$IDF_t = \log \frac{N}{DF_t} \tag{4}$$

$W_{t,d}$ adalah nilai yang mewakili pentingnya kata t dalam dokumen d, $TF_{t,d}$ mengacu pada seberapa sering kata t muncul dalam dokumen d, dan IDF_t menggambarkan seberapa umumnya kata t dalam semua dokumen, dengan N sebagai total dokumen yang ada, dibagi dengan DF_t yang merupakan jumlah dokumen yang mengandung kata term t.

4. SMOTE

SMOTE (Synthetic Minority Oversampling Technique) merupakan sebuah teknik yang dipakai untuk menngatasi ketidakseimbangan kelas. Teknik ini bekerja untuk menambahkan data baru (sintesis) pada data minoritas agar jumlah data seimbang dengan data mayoritas [16]. Ini menghindarkan model machine learning dari kenderungan untuk lebih memperhatikan kelas mayoritas karena jumlahnya yang lebih besar. Dengan menggunakan SMOTE, model dapat belajar dari kelas minoritas dengan lebih baik, meningkatkan untuk mengklasifikasi kedua kelas dengan akurat

5. Split Dataset

Selama langkah pemisahan data, himpunan data dipartisi menjadi data pelatihan (80%) dan data pengujian (20%) dalam metode ini. Data training dipakai untuk melatih agar algoritma dapat mengidentifikasi yang mana kelas positif dan kelas negatif dalam dataset. Setelah melatih algoritma dengan data training, selanjutnya adalah menguji kinerja algoritma dengan menggunakan data testing.

6. Modelling

Pada langkah ini akan dilaksanakan pemodelan klasifikasi pada data ulasan yang sebelumnya telah dilakukan preprocessing, feature extraction, smote, dan split data. Kemudian penulis menerapkan model klasifikasi yang dipakai ialah algoritma Naïve Bayes dan SVM untuk memperoleh nilai akurasi pada model tersebut yang nantinya akan dikomparasikan. Pada kedua algoritma tersebut peneliti menggunakan teknik SMOTE untuk menangani ketidakseimbangan kelas.

7. Evaluasi

Evaluasi model dilakukan penelitian melalui confusion matrix. Confusion matrix ialah sebuah metode yang biasa dipakai untuk menilai sejauh mana kinerja suatu model dalam proses mengklasifikasikan dataset [17]. Dengan matriks ini, menggambarkan tabel 2x2 yang

memberikan rincian terkait hasil klasifikasi dalam meprediksi kelas yang benar dan seberapa sering model melakukan kesalahan dalam prediksi kelas. Dalam evaluasi kinerja terdapat empat elemen yaitu, TP jumlah data testing dengan katagori positif diprediksi benar, TN jumlah data testing dengan katagori negatif diprediksi benar, FN jumlah data testing dengan katagori Negatif diprediksi salah, FP Jumlah data testing dengan katagori positif diprediksi salah. Confusion matrix bisa diliat pada tabel 2.

TABEL III
COFUISION MATRIX

Aktual	Prediksi Negatif	Prediksi Positif
Negatif	*True Negative (TN*)	*False Negative (FN)*
Positif	*False Positive (FP)*	*True Positive (TP)*

Pada Tabel 2, Dengan infromasi yang diberikan confusion matrix dapat menghitung rumus metrik yang menghasilkan kinerja suatu model seperti precision, accuracy, recall, dan f1-score. Rumus accuracy, precision, recall, dan f1-score yaitu:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{f1-score} = \frac{2TP}{2TP+FP+FN} \quad (8)$$

8. Visualisasi Wordcloud

Proses visualisasi akan menampilkan visualisassi wordcloud. Wordcloud adalah visualisasi data yang menampilkan fitur teks. Visualisasi ini akan menampilakan kumpulan kata-kata dengan berabagai ukuran, semakin tinggi kumpulan kata dalam wordcloud, semakin banyak pula kata yang mucul dalam teks yang dianalsis. Hal ini memudahkan untuk melihat kata-kata yang paling sering dipakai dan memahami topik utama dalam teks [18].

III. HASIL DAN PEMBAHASAN

A. Dataset

Dataset didapatkan dari review di google play store dengan proses web scrapping sebanyak 3000 dataset dengan retang waktu dari september 2023 hingga febuari 2024. Hasil dari scrapping ditunjukkan pada gambar 3.

	content	score
0	Sebenarnya capcut ini udah cukup membantu sih ...	3
1	Mantap si ini apk yang memiliki fitur komplit ...	4
2	Bagusih.. Tpi update nya bisa diperbaiki kaya ...	4
3	Oke bagus Tapi Bug nya !!.... Semakin hari/sem...	4
4	apk nya udh bagus bgt dan ngebantu buat ngebik...	3
...	...	...
2995	Bagus bangettt....bisa ngedit video2 dan ngiri...	5
2996	Gak tau kenpa . Hbis update .setiap mau cari t...	4
2997	Pencarian masih ga bisa tuh boss, sinyak kence...	1
2998	ngelek banget pas buka apk nya trs juga sering...	1
2999	aplikasi ini sangat bagus , sangat membantu , ...	5

3000 rows × 2 columns

Gambar 3. Hasil scrapping

B. Preprocessing

1. Cleansing

Cleansing merupakan tahap awal proses pengolahan data dengan tujuan untuk membersihkan karakter dalam data text seperti, tanda baca, menghilangkan emotion, url (link), dan hastag. Hasil cleansing bisa dipakai pada tabel 3.

TABEL IIIII
HASIL PROCESSING CLEANSING

Sebelum	**Sesudah**
Mantap. Sukses selalu untuk capcut semoga memberikan fitur fitur yang terbaru lagi yang memudahkan para konten kreator nya serta membuat semangat untuk membuat kreasi berbagai konten video. Semoga saya bisa cepet di kasih cuan edit template capcut. Mudah dipahami dan simpel untuk segi edit buat konten videonya. 😎🤤	Mantap Sukses selalu untuk capcut semoga fitur fitur yang terbaru lagi yang para konten kreator nya serta membuat semangat untuk membuat kreasi berbagai konten video Semoga saya bisa cepet di kasih cuan edit template capcut Mudah dipahami dan simpel untuk segi edit buat konten videonya

2. Case Folding

Case folding ialah proses yang bertujuan untuk mengonversi semua huruf dalam teks jadi huruf kecil secara menyeluruh [19]. Ssperti kata “Mantap” menjadi “mantap” hal ini yang ditunjukkan pada tabel 4 hasil dari *case folding*.

TABEL IVV
HASIL PROCESSING CACE FOLDING

Sebelum	Sesudah
Mantap Sukses selalu untuk capcut semoga fitur fitur yang terbaru lagi yang para konten kreator nya serta membuat semangat untuk membuat kreasi berbagai konten video Semoga saya bisa cepet di kasih cuan edit template capcut Mudah dipahami dan simpel untuk segi edit buat konten videonya	mantap sukses selalu untuk capcut semoga fitur fitur yang terbaru lagi yang para konten kreator nya serta membuat semangat untuk membuat kreasi berbagai konten video semoga saya bisa cepet di kasih cuan edit template capcut mudah dipahami dan simpel untuk segi edit buat konten videonya

3. Tokenizing

Tokenizing adalah proses memisahkan kata perkata yang membentuk suatu kalimat, untuk membedakan polaritas sentimen setiap kalimat dengan lebih mudah [20]. Hasil dapat diliat pada tabel 5.

TABEL V
HASIL PROCESSING TOKENIZING

Sebelum	Sesudah
mantap sukses selalu untuk capcut semoga fitur fitur yang terbaru lagi yang para konten kreator nya serta membuat semangat untuk membuat kreasi berbagai konten video semoga saya bisa cepet di kasih cuan edit template capcut mudah dipahami dan simpel untuk segi edit buat konten videonya	'mantap', 'sukses', 'selalu', 'untuk', 'capcut', 'semoga', 'fitur', 'fitur', 'yang', 'terbaru', 'lagi', 'yang', 'para', 'konten', 'kreator', 'nya', 'serta', 'membuat', 'semangat', 'untuk', 'membuat', 'kreasi', 'berbagai', 'konten', 'video', 'semoga', 'saya', 'bisa', 'cepet', 'di', 'kasih', 'cuan', 'edit', 'template', 'capcut', 'mudah', 'dipahami', 'dan', 'simpel', 'untuk', 'segi', 'edit', 'buat', 'konten', 'videonya'

4. Filtering (Stopword Removal)

Filtering atau *stopword removal* merupakan proses yang digunakan untuk menghapus kata-kata yang seringkali ada pada dokumen tetapi tidak mengandung informasi yang dipehitungkan dalam analisis sentimen. Tujuannya adalah untuk mempersempit dimensi data, menaikan efesiensi komputasi, dan memfokuskan pada kata-kata yang lebih penting untuk dianalisis [21]. Hasil stopword removal bisa ditinjau pada tabel 6.

TABEL VI
HASIL PROCESSING FILTERING

Sebelum	Sesudah
'mantap', 'sukses', 'selalu', 'untuk', 'capcut', 'semoga', 'fitur', 'fitur', 'yang', 'terbaru', 'lagi', 'yang', 'para', 'konten', 'kreator', 'nya', 'serta', 'membuat', 'semangat', 'untuk', 'membuat', 'kreasi', 'berbagai', 'konten', 'video', 'semoga', 'saya', 'bisa', 'cepet', 'di', 'kasih', 'cuan', 'edit', 'template', 'capcut', 'mudah', 'dipahami', 'dan', 'simpel', 'untuk', 'segi', 'edit', 'buat', 'konten', 'videonya'	'mantap', 'sukses', 'capcut', 'semoga', 'fitur', 'fitur', 'yang', 'terbaru', 'yang', 'konten', 'kreator', 'nya', 'semangat', 'kreasi', 'konten', 'video', 'semoga', 'cepet', 'kasih', 'cuan', 'edit', 'template', 'capcut', 'mudah', 'dipahami', 'simpel', 'segi', 'edit', 'konten', 'videonya'

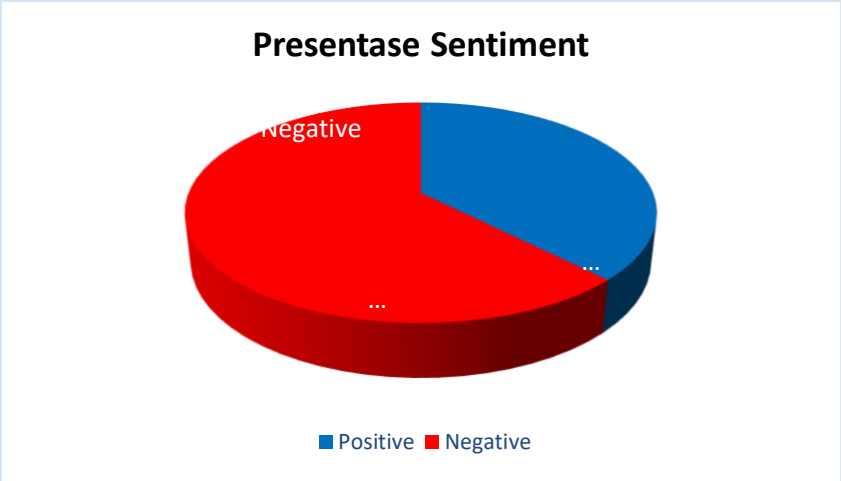
5. Stemming

Stemming merupakan proses yang merubah kata-kata jadi bentuk dasar atau kata dasar [22]. Maksudnya ialah untuk menyederhanakan kata-kata dalam sebuah teks dengan menghaspus imbuhan sehingga menemukan kata dasar yang memiliki nilai untuk mewakili makna keseluruhan dari kata.

TABEL VII
HASIL PROCESSING STEMMING

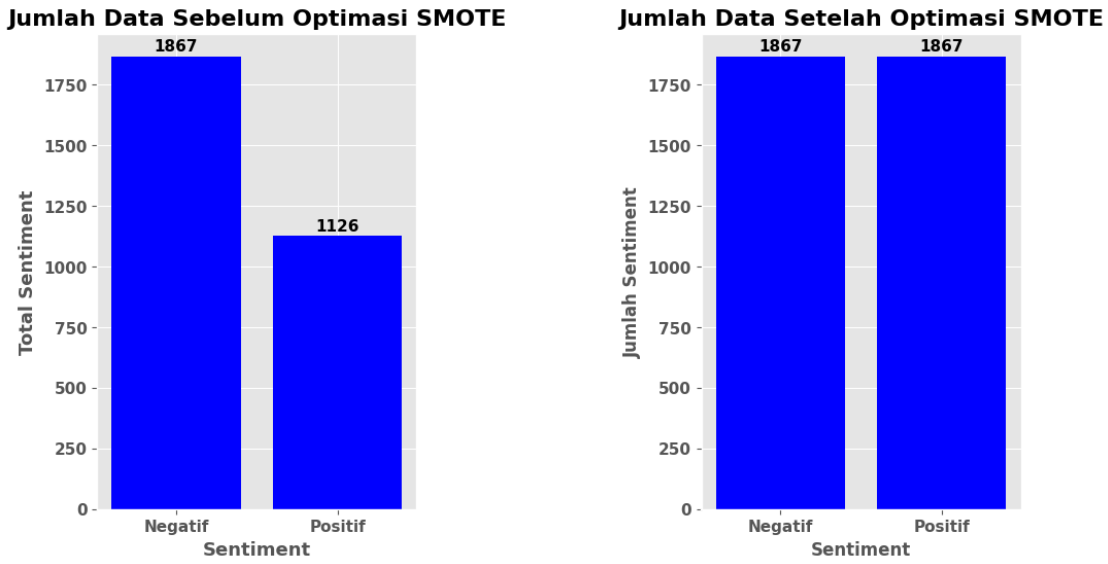
Sebelum	Sesudah
'mantap', 'sukses', 'capcut', 'semoga', 'fitur', 'fitur', 'yang', 'terbaru', 'yang', 'konten', 'kreator', 'nya', 'semangat', 'kreasi', 'konten', 'video', 'semoga', 'cepet', 'kasih', 'cuan', 'edit', 'template', 'capcut', 'mudah', 'dipahami', 'simpel', 'segi', 'edit', 'konten', 'videonya'	mantap sukses capcut moga fitur fitur yang baru yang konten kreator nya semangat kreasi konten video moga cepet kasih cuan edit template capcut mudah paham simpel segi edit konten video

Seletah melewati tahap preprocessing, jumlah data ulasan yang awalnya 3000 menjadi 2993 dataset. Selanjutnya akan dilakukan pelabelan data berdasarkan rating ulasan yang dimana rating 1 – 3 dianggap sentimen negatif dan rating 4 – 5 dianggap sentimen positif. Sentimen negatif mendapatkan sebanyak 1867 ulasan, sedangkan sentimen positif mendapatkan 1126 ulasan. Nilai presentase sentimen ulasan pengguna capcut ditunjukkan pada gambar 4.



Gambar 4. Presentase sentimen ulasan pengguna capcut

Dalam nilai presentase sentimen, terjadi ketidakseimbangan data antara sentimen negatif dan positif. Ketika mayoritas mendominasi dataset, hal ini dapat menyebabkan model akan lebih cenderung melatih data pada negatif. sehingga performa model dalam memprediksi sentimen positif kurang optimal. Maka dari itu pada penelitian ini diterapkan teknik *SMOTE* untuk mengatasi ketidakseimbangan antar kelas. Perbandingan sebelum dan setelah memakai *SMOTE* bisa ditinjau pada gambar 5.



Gambar 5. Jumlah Sentimen setelah memakai *SMOTE*

Pada data kelas minor, akan ditambahkan data sintetis pada sentimen positif sebanyak 741 data untuk menyamakan data pada kelas mayor yaitu sentimen negatif, agar model algoritma akan lebih seimbang untuk mempelajari suatu sentimen negatif maupun positif. Setelah optimasi SMOTE jumlah pada sentimen negatif dan posistif sebanyak 1867 data.

C. Evaluasi Algoritma

Setelah melalui beberapa tahap persiapan data, hasil pengujian akan dilakukan dengan komparasi algoritma Naïve Bayes dan SVM dengan mengunakan data training senilai 80% dan data testing senilai 20%. Berdasarkan data yang sudah terkumpul dan siap digunakan untuk analisis, data ini dievaluasi dalam dua kondisi, yaitu sebelum dan setelah diterapkannya teknik

SMOTE. Pada tabel 8 akan memperilatkan nilai precision, accuracy, recall, f1-score sebagai parameter perbandingan.

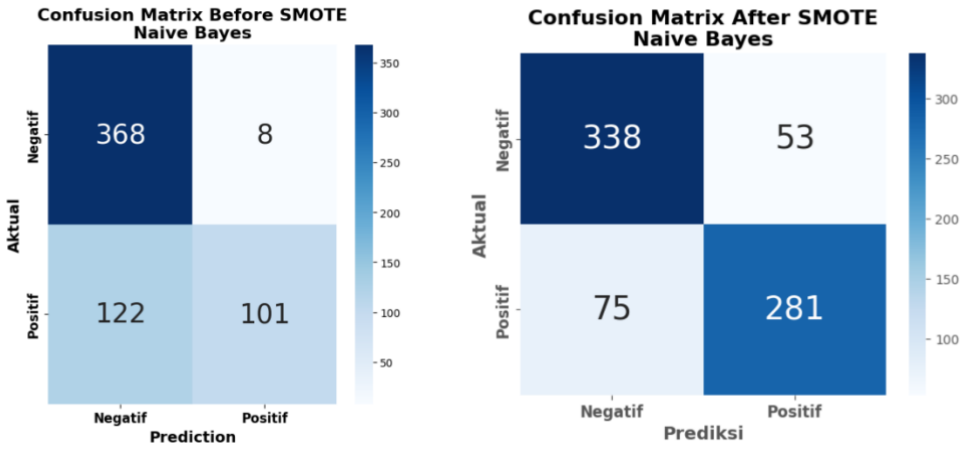
TABEL VIII
PERBANDINGAN ALGORITMA SEBELUM DAN SESUDAH *SMOTE*

Matriks	NB	NB + SMOTE	SVM	SVM + SMOTE
Accuracy	78%	83%	81%	86%
Sentimen Negatif				
Precision	75%	82%	81%	86%
Recall	98%	86%	91%	87%
F1-Score	85%	84%	86%	86%
Sentimen Positif				
Precision	93%	84%	82%	85%
Recall	45%	79%	63%	85%
F1-Score	61%	81%	71%	85%

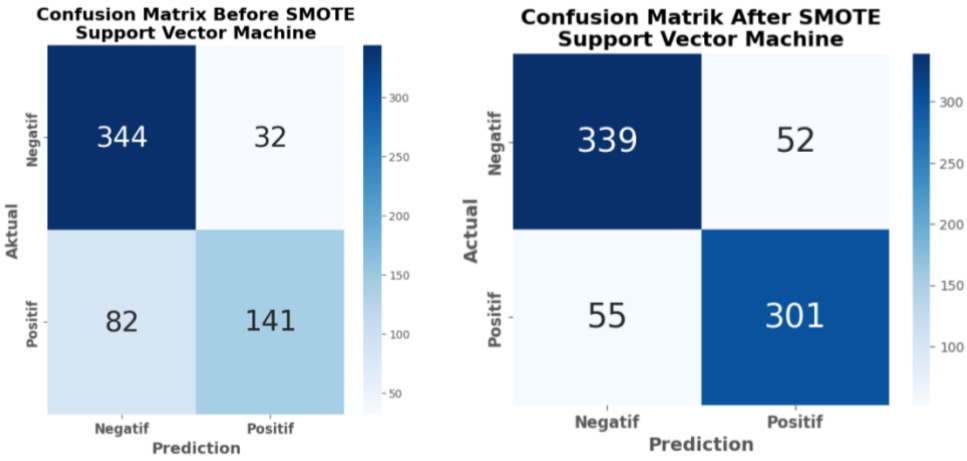
Berdasarkan pada tabel 8, menunjukan bahwa model Niave Bayes dan SVM menghasilkan akurasi yang cukup baik. Sebelum menggunakan teknik *SMOTE* model Naïve Bayes memperoleh akurasi 78% sedangkan model SVM memperoleh akurasi 81%. Setelah memperoleh teknik *SMOTE*, akurasi model Naïve Bayes meningkat menjadi 83% sedangkan akurasi model SVM juga meningkat menjadi 86%. Dari masing-masing model, terjadi peningkatan nilai accuracy sebanyak 5% setelah menerapkan teknik *SMOTE*.

Adapun nilai precision, recall, f1-score yang dapat memberikan gambaran yang lengkap terkait kinerja suatu model klasifikasi. Hasil pengujian model Naïve Bayes dengan *SMOTE*, terdapat sentimen positif pada nilai recall dan f1-score mengalami peningkatan. Nilai recall meningkat dari 45% menjadi 79%, dan nilai f1-score meningkat dari 61% menjadi 81%. Namun peningkatan ini diiringi dengan penurunan pada precision. Sedangkan hasil pengujian model SVM dengan *SMOTE* juga mengalami peningkatan, terdapat sentimen positif pada nilai precision, recall, dan f1-score mengalami peningkatan, yang dimana nilai precesion meningkat dari 82% menjadi 85%, nilai recall meningkat 63% menjadi 85%, dan nilai f1-score meningkat dari 71% menjadi 85%.

Penelitian ini juga akan melakukan komparasi nilai confusion matrix dengan memahami sejauh mana kedua algoritma berhasil mengklasifikasikan data dengan benar dan mengidetifikasi kelasahan yang terjadi. Pada gambar 6 dan 7 menunjukan hasil cofusion matrix setiap model.



Gamber 6. *Confusion matrix* model Naïve Bayes before dan after *SMOTE*



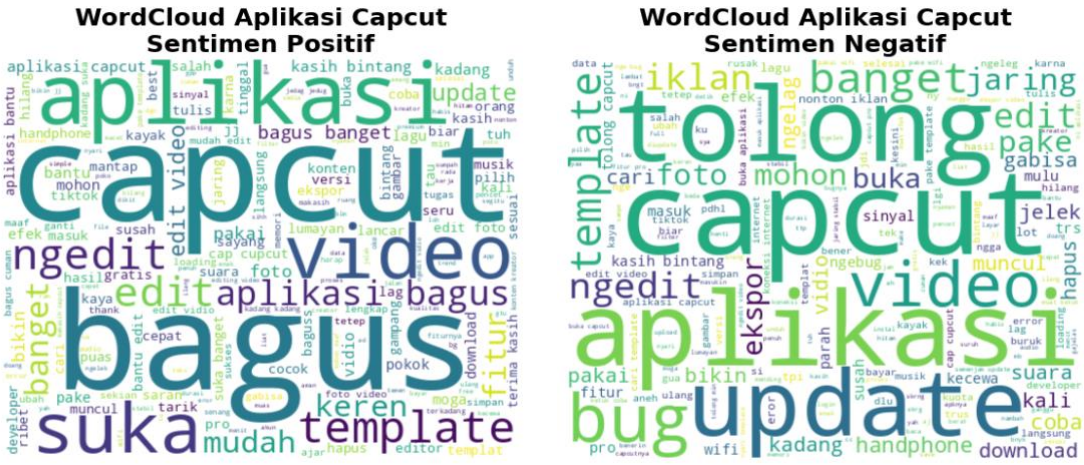
Gambar 7. *Confusion matrix* model SVM before dan after *SMOTE*

Hasil dari confusion matrix sebelum menggunakan *SMOTE*, dari kedua algoritma *Niave Bayes* dan SVM menunjukkan ketidakmampuannya dalam membedakan kata positif secara opitimal. Hal ini dikarnakan model lebih terfokus pada pembelajaran kata negatif yang merupakan kelas mayoritas. Akibatnya banyak data positif yang terklasifikasi keliru sebagai data negatif. Namun, setelah optimasi *SMOTE* diterapkan, model mulai bisa mengenali kata positif secara optimal terlihat dari peningkatan *(True Positive)* pada algoritma Naïve Bayes dari 101 data menjadi 281 data, dan pada algoritma SVM juga meningkat dari 141 data menjadi 301 data.

Berdasarkan data yang diberikan dengan confusion matrix dapat simpulkan bahwa SVM *(Support Vector Machine)* terbukti sebagai algoritma terbaik. Hal ini didasari dengan komparasi algoritma Naïve Bayes dan SVM menggunakan optimasi SMOTE. Algoritma SVM menunujukkan akurasi 86% lebih unggul dibandingkan Naïve Bayes dengan akurasi 81%. Dan SVM juga terbukti lebih akurat dalam mengenali kategori yang ingin diklasifikasikan dibandingkan Niave Bayes. Hal tersebut ditunjukkan oleh peningkatan confusion matrix pada nilai TP *(True Positive).*

D. Hasil Visualisasi Data

Pada tahap hasil visualisasi memakai wordcloud yang akan menunjukan sebaran kata-kata yang paling sering mucul dalam review pemakai pada aplikasi capcut yang bersentimen positif atau negatif yang dipaparkan pada gambar 9.



Gambar 8. Wordcloud aplikasi capcut pada sentimen positif dan negatif

Pada gambar 8, hasil dari visualisasi *wordcloud* pada sentimen positif terdapat frekuensi kata yang sering muncul seperti, “capcut”, “aplikasi”, “bagus”, “video”, “template”, “suka”, dan lainya. Penggunaan kata “capcut”, ”aplikasi”, “bagus”, “suka”, merujuk pada pengguna menyukai aplikasi capcut dan merasa puas untuk mengedit video. Penggunaan kata “template” dan “video” merujuk pada fitur template yang disediakan telah membantu pengguna memudahkan proses pengeditan video.

Disisi lain *wordcloud* pada sentimen negatif, kata-kata yang sering muncul antara lain, “aplikasi”, “tolong”, “update”, “bug”, “iklan”, “video”, “template”, “ekspor”, dan lainnya. Pengunaan kata “update” dan “bug” merujuk pada adanya masalah teknis atau kesalahan dalam pembaruan aplikasi yang mengganggu pengguna. Selain itu penggunaan kata “iklan”, “video”, “template” dan “ekspor” merujuk ada masalah atau kediaknyamanan yang dialami pengguna ketika sedang menggunakan aplikasi capcut, sepeti iklan yang menggaggu, masalah dalam pengeditan video atau template, atau kendala proses ekspor hasil editing. Demikian pula, permintaan bantuan atau perbaikan yang disampaikan oleh pengguna melalui kata “tolong” juga harus dipertimbangkan.

IV. KESIMPULAN

Berdasarkan hasil penelitian terhadap 2993 data ulasan aplikasi CapCut di Google Play Store, ditemukan bahwa 62% ulasan memiliki sentimen negatif (1867 data), sedangkan 38% ulasan memiliki sentimen positif (1126 data). Penelitian ini dilakukan komparasi algoritma Naive Bayes dan SVM dalam analisis sentimen ulasan tersebut. Hasil komparasi menunujukan bahwa Naïve Bayes memiliki nilai akurasi sebesar 78%, precision 93%, recall 45%, f1-score 61%, dan Support Vector Machine mendapatkan 81% , precision 82%, recall 63%, f1-score 71%. Berdasarkan eksperimen yang telah dilakukan model terebut masih mendominasi belajar lebih banyak data sentimen negatif sehingga terjadi ketidakseimbangan data. Oleh karena itu peneliti melakukan optimasi SMOTE agar kelas minoritas dapat memiliki jumlah yang sama dengan kelas mayoritas. Dengan menggunakan metode SMOTE, khususnya dalam mengenali sentimen positif, model SVM mengalami kenaikan dibandingkan dengan Naïve Bayes yang terjadi penurunan pada precision, algoritma SVM mendapatkan nilai akurasi 86%, precision 85%, recall 85%, f1-score 85%. Berdasarkan visualisasi *wordcloud* pada ulasan positif menunjukkan bahwa pengguna menyukai aplikasi Capcut dan merasa puas dengan fitur-fiturnya, terutama kemudahan dalam mengedit video dan template yang disediakan. Sedangkan pada ulasan negatif menunjukkan bahwa pengguna mengalami masalah teknis atau kesalahan dalam pembaruan aplikasi, iklan yang mengganggu, terjadinya bug saat proses pengeditan, dan masalah dalam proses ekspor hasil editing. Oleh karena itu pengembang harus fokus pada pengujian yang lebih ketat sebelum merilis pembaruan dan memperbaiki bug dengan cepat untuk mengatasi masalah teknis. Pengelolaan iklan yang lebih baik diperlukan untuk mengurangi gangguan, misalnya dengan menyediakan opsi premium tanpa iklan. Selain itu, memastikan bahwa fitur ekspor berfungsi dengan baik untuk meningkatkan pengalaman pengguna. Penelitian ini memiliki dalam menggunakan dua algoritma yaitu Naïve Bayes dan Support Vector Machine (SVM). Maka dari itu peneliti sarankan menggunakan model klasifikasi lain yang lebih beragam untuk dapat meningkatkan akurasi yang lebih tinggi.

REFERENSI

[1] A. Tedyyana, O. Ghazali, and O. W. Purbo, “Machine learning for network defense: automated DDoS detection with telegram notification integration,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 2, p. 1102, May 2024, doi: 10.11591/ijeecs.v34.i2.pp1102-1109.

[2] Hetilaniar, F. Rokhman, and R. Pristiwati, “Dari Dunia Offline ke Dunia Online: Merangkul Literasi Digital,” *Jurnal Pembahsi (Pembelajaran Bahasa Dan Sastra Indonesia)*, vol. 13, no. 1, pp. 44–54, 2023, doi: 10.31851/pembahsi.v13i1.11936.

[3] Friska Aditia Indriyani, Ahmad Fauzi, and Sutan Faisal, "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan support vector machine," *TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika*, vol. 10, no. 2, pp. 176–184, 2023, doi: 10.37373/tekno.v10i2.419.

[4] A. Tedyyana and O. Ghazali, "Teler Real-time HTTP Intrusion Detection at Website with Nginx Web Server," *JOIV : International Journal on Informatics Visualization*, vol. 5, no. 3, p. 327, Sep. 2021, doi: 10.30630/joiv.5.3.510.

[5] A. Fricticarani, A. Hayati, R. R, I. Hoirunisa, and G. M. Rosdalina, "Strategi Pendidikan Untuk Sukses Di Era Teknologi 5.0," *Jurnal Inovasi Pendidikan dan Teknologi Informasi (JIPTI)*, vol. 4, no. 1, pp. 56–68, 2023, doi: 10.52060/pti.v4i1.1173.

[6] E. Sutisna, F. Angellia, I. Pranawukir, and E. Efendi, "Analisis Pengaruh Penggunaan Aplikasi Capcut Terhadap Keterlibatan Dan Kesetiaan Pelanggan," *Journal of Computer Science and Information Technology*, vol. 1, no. 1, pp. 27–34, 2023, doi: 10.59407/jcsit.v1i1.333.

[7] R. Syahmewah, "Pengaruh Penggunaan Template Pada Aplikasi Capcut Yang Memudahkan Mahasiswa Untuk Mengedit Video Sebagai Media Pembelajaran," *Journal of Physics and Science Learning*, vol. 07, no. 1, pp. 27–32, 2023.

[8] M. K. Khoirul Insan, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 478–483, 2023, doi: 10.36040/jati.v7i1.6373.

[9] A. M. Ndapamuri, D. Manongga, and A. Iriani, "Analisis Sentimen Ulasan Aplikasi Tripadvisor Dengan Metode Support Vector Machine, K-Nearest Neighbor, Dan Naive Bayes," *INOVTEK Polbeng - Seri Informatika*, vol. 8, no. 1, p. 127, 2023, doi: 10.35314/isi.v8i1.3260.

[10] A. Mustofa and R. Novita, "Klasifikasi Sentimen Masyarakat Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat Menggunakan Text Mining Pada Twitter," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 200–208, 2022, doi: 10.47065/bits.v4i1.1628.

[11] M. Alfarizi, M. Rizqy, R. I. Ghufroni, D. Fathurahman, R. D. Saputra, and F. Kurniawan, "Analisis Sentimen Persepsi Publik Terhadap Kasus Bullying Siswa Cilacap Menggunakan Pendekatan Machine Learning," *Journal of Information Technology Ampera*, vol. 4, no. 3, pp. 265–276, 2023.

[12] M. I. Ghozali, W. H. Sugiharto, and A. F. Iskandar, "Analisis Sentimen Pinjaman Online Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 6, pp. 1340–1348, 2023, doi: 10.30865/klik.v3i6.936.

[13] Ernianti Hasibuan and Elmo Allistair Heriyanto, "Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier," *Jurnal Teknik dan Science*, vol. 1, no. 3, pp. 13–24, 2022, doi: 10.56127/jts.v1i3.434.

[14] V. W. D. Thomas and F. Rumaisa, "Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, p. 1767, 2022, doi: 10.30865/mib.v6i3.4218.

[15] K. Pramayasa, I. M. D. Maysanjaya, and I. G. A. A. D. Indradewi, "Analisis Sentimen Program Mbkm Pada Media Sosial Twitter Menggunakan KNN Dan SMOTE," *SINTECH (Science and Information Technology) Journal*, vol. 6, no. 2, pp. 89–98, 2023, doi: 10.31598/sintechjournal.v6i2.1372.

[16] R. Fatiya *et al.*, "Pengaruh Synthetic Minority Oversampling Technique pada Analisis Sentimen Menggunakan Algoritma K-Nearest Neighbors," *Jlk*, vol. 5, no. 1, pp. 7–12, 2022.

[17] D. Prasetyawan and R. Gatra, "Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 7, no. 1, pp. 56–67, 2022, doi: 10.14421/jiska.2022.7.1.56-67.

[18] D. T. Lukmana, S. Subanti, and Y. Susanti, "Analisis Sentimen Terhadap Calon Presiden 2019 Dengan Support Vector Machine Di Twitter," *Seminar Nasional Penelitian Pendidikan Matematika (SNP2M) 2019 UMT*, no. 2002, pp. 154–160, 2019.

[19] 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.

[20] T. Mardiana, H. Syahreva, and T. Tuslaela, "Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, 2019, doi: 10.33480/pilar.v15i2.752.

[21] Fauzan Baehaqi and N. Cahyono, "Analisis Sentimen Terhadap Cyberbullying Pada Komentar Di Instagram Menggunakan Algoritma Naïve Bayes," *Indonesian Journal of Computer Science*, vol. 13, no. 1, pp. 1051–1063, 2024, doi: 10.33022/ijcs.v13i1.3301.

[22] C. Prakoso and A. Hermawan, "Perbandingan Model Machine Learning dalam Analisis Sentimen Ulasan Pengunjung Keraton Yogyakarta pada Google Maps," *Media Online*, vol. 4, no. 3, pp. 1292–1302, 2023, doi: 10.30865/klik.v4i3.1419.