ANALISIS SENTIMEN CYBERBULLYING PADA KOMENTAR TIKTOK TERKAIT ISU REMAJA MENGGUNAKAN NAÏVE BAYES CLASSIFIER SEBAGAI DASAR PENGUATAN LITERASI DIGITAL

M. Saleh[1], M. Akbar Riwanto[2], Muhammad Ari Ardana[3], Muh. Rasyid Ridha[4]
[1-4]Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Islam Indragiri,
Email: msaleh0205@gmail.com[1], akbarrwnt@gmail.com[2], ardana1520@gmail.com[3], rasyid4sky@gmail.com[4]

ABSTRAK

Cyberbullying merupakan salah satu permasalahan yang sering muncul pada media sosial dan dapat berdampak negatif terhadap kesehatan mental remaja. TikTok sebagai platform yang banyak digunakan remaja memungkinkan terjadinya interaksi melalui komentar yang mengandung sentimen positif maupun negatif. Penelitian ini bertujuan untuk menganalisis sentimen komentar TikTok terkait isu remaja menggunakan algoritma Multinomial Naïve Bayes sebagai dasar penguatan literasi digital. Penelitian menggunakan pendekatan kuantitatif dengan data sebanyak 1.508 komentar TikTok. Tahapan penelitian meliputi pengumpulan data, pelabelan, preprocessing teks (*case folding, cleaning, tokenizing, stopword removal,* dan *stemming*), transformasi data menggunakan TF-IDF, serta klasifikasi menggunakan Multinomial Naïve Bayes. Evaluasi model dilakukan menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa 55,2% komentar tergolong sentimen positif dan 44,8% sentimen negatif. Model menghasilkan nilai *accuracy* sebesar 69,33%, *precision* sebesar 68,14%, *recall* sebesar 83,73%, dan *F1-score* sebesar 75,14%. Hasil tersebut menunjukkan bahwa Multinomial Naïve Bayes memiliki kemampuan yang cukup baik dalam mengklasifikasikan sentimen komentar TikTok dan mengidentifikasi potensi cyberbullying. Temuan penelitian ini dapat dimanfaatkan sebagai dasar pengembangan program literasi digital dan upaya pencegahan cyberbullying pada kalangan remaja.

Kata Kunci: Analisis Sentimen, Cyberbullying, TikTok, Naïve Bayes, Literasi Digital

ABSTRACT

Cyberbullying is a common problem on social media that can have a negative impact on adolescents' mental health. TikTok, a platform widely used by adolescents, facilitates interactions through comments that contain both positive and negative sentiments. This study aims to analyze the sentiment of TikTok comments related to adolescent issues using the Multinomial Naïve Bayes algorithm as a basis for strengthening digital literacy. The study employed a quantitative approach using a dataset of 1,508 TikTok comments. The research stages included data collection, labeling, text preprocessing (case folding, cleaning, tokenization, stopword removal, and stemming), data transformation using TF-IDF, and classification using the Multinomial Naïve Bayes algorithm. Model evaluation was conducted using a confusion matrix, accuracy, precision, recall, and F1-score. The results showed that 55.2% of the comments were classified as positive sentiment and 44.8% as negative sentiment. The model achieved an accuracy of 69.33%, a precision of 68.14%, a recall of 83.73%, and an F1-score of 75.14%. These results indicate that the Multinomial Naïve Bayes model performs quite well in classifying the sentiment of TikTok comments and identifying potential cyberbullying. The findings of this study can be used as a basis for developing digital literacy programs and efforts to prevent cyberbullying among adolescents.

Translated with DeepL.com (free version)

Keywords: Sentiment Analysis, Cyberbullying, TikTok, Naïve Bayes, Digital Literacy

1 PENDAHULUAN

Perkembangan teknologi digital dan internet telah mengubah cara masyarakat berkomunikasi, memperoleh informasi, serta berinteraksi dalam kehidupan sehari-hari. Media sosial menjadi salah satu produk perkembangan teknologi yang mengalami pertumbuhan sangat pesat dan telah menjadi bagian penting dari kehidupan generasi muda. Berdasarkan laporan *Digital 2024: Global Overview Report*, jumlah pengguna media sosial dunia terus mengalami peningkatan dan telah mencapai lebih dari lima miliar pengguna, dengan kelompok usia remaja dan dewasa muda sebagai pengguna paling aktif [1]. Di Indonesia, survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menunjukkan bahwa penetrasi internet terus meningkat dan didominasi oleh kelompok usia produktif, termasuk remaja yang menggunakan internet untuk berkomunikasi, mencari informasi, hiburan, serta berinteraksi melalui berbagai platform media sosial [2]. Kondisi ini menunjukkan bahwa media sosial telah menjadi ruang komunikasi digital yang memiliki pengaruh besar terhadap perilaku dan pola interaksi remaja.

Salah satu platform media sosial yang mengalami pertumbuhan pengguna paling pesat dalam beberapa tahun terakhir adalah TikTok. Platform berbasis video pendek ini memungkinkan pengguna untuk membuat, membagikan, serta memberikan tanggapan terhadap konten melalui fitur komentar secara real-time. Karakteristik TikTok yang mengedepankan interaksi cepat, kemudahan berbagi konten, serta algoritma rekomendasi yang personal menjadikan platform ini sangat populer di kalangan remaja. Selain memberikan manfaat berupa sarana ekspresi diri, kreativitas, hiburan, dan pembelajaran, TikTok juga menciptakan ruang komunikasi yang sangat terbuka sehingga memungkinkan munculnya berbagai bentuk interaksi negatif antar pengguna [3], [4]. Tingginya intensitas komunikasi yang terjadi pada kolom komentar menjadikan TikTok sebagai salah satu platform yang berpotensi menjadi media terjadinya perilaku agresif secara daring.

Salah satu bentuk interaksi negatif yang banyak ditemukan pada media sosial adalah *cyberbullying*. Cyberbullying merupakan tindakan agresif yang dilakukan secara sengaja dan berulang melalui media digital dengan tujuan menyakiti, mempermalukan, mengintimidasi, atau mengancam individu lain [5]. Perilaku ini dapat muncul dalam berbagai bentuk, seperti penghinaan, ujaran kebencian, pelecehan verbal, *body shaming*, penyebaran informasi yang merugikan, maupun komentar yang bersifat merendahkan [6]. Berbeda dengan perundungan konvensional, cyberbullying memiliki jangkauan yang lebih luas karena dapat dilakukan kapan saja dan menyebar dengan cepat kepada banyak pengguna dalam waktu singkat. Akibatnya, dampak yang ditimbulkan juga dapat berlangsung lebih lama dan memengaruhi kondisi psikologis korban secara signifikan.

Berbagai penelitian menunjukkan bahwa cyberbullying memiliki dampak negatif terhadap kesehatan mental remaja. Korban cyberbullying berisiko mengalami penurunan kepercayaan diri, kecemasan, stres, depresi, gangguan emosional, hingga penurunan kualitas hidup [7]. UNICEF dalam *The State of the World's Children 2024* menegaskan bahwa remaja merupakan kelompok yang paling rentan terhadap risiko interaksi digital yang tidak sehat, termasuk perundungan daring dan berbagai bentuk kekerasan berbasis teknologi [8]. Tingginya penggunaan media sosial oleh remaja menjadikan kelompok ini semakin rentan terhadap paparan komentar negatif yang berpotensi mengganggu perkembangan sosial maupun psikologis mereka. Oleh karena itu, diperlukan upaya untuk mengidentifikasi dan meminimalkan keberadaan cyberbullying pada media sosial sebagai langkah preventif dalam menciptakan lingkungan digital yang lebih aman.

Salah satu pendekatan yang banyak digunakan untuk mengidentifikasi kecenderungan opini pengguna media sosial adalah analisis sentimen (*sentiment analysis*). Analisis sentimen merupakan cabang dari *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau emosi yang terkandung dalam suatu teks ke

dalam kategori tertentu, seperti sentimen positif, negatif, atau netral [9], [10]. Pada media sosial, analisis sentimen tidak hanya digunakan untuk memahami opini publik terhadap suatu isu, tetapi juga dimanfaatkan untuk mendeteksi komentar yang mengandung ujaran kebencian, perilaku agresif, maupun indikasi cyberbullying [11]. Dengan menganalisis sentimen yang terkandung dalam komentar pengguna, peneliti dapat memperoleh gambaran mengenai pola komunikasi yang terjadi dalam suatu komunitas digital serta mengidentifikasi potensi risiko yang muncul dari interaksi tersebut.

Keberhasilan analisis sentimen sangat dipengaruhi oleh metode klasifikasi yang digunakan. Salah satu algoritma yang banyak diterapkan dalam klasifikasi teks adalah Multinomial Naïve Bayes. Algoritma ini merupakan metode klasifikasi probabilistik yang bekerja berdasarkan Teorema Bayes dengan asumsi independensi antar fitur [12]. Multinomial Naïve Bayes memiliki beberapa keunggulan, antara lain sederhana, efisien dalam proses komputasi, mampu menangani data berdimensi tinggi, serta memberikan performa yang cukup baik pada klasifikasi teks [12], [13]. Beberapa penelitian menunjukkan bahwa kombinasi pembobotan TF-IDF dan Multinomial Naïve Bayes mampu menghasilkan akurasi yang kompetitif dalam analisis sentimen berbahasa Indonesia [13], [14]. Oleh karena itu, algoritma ini masih banyak digunakan dalam penelitian yang berkaitan dengan analisis sentimen dan klasifikasi teks pada media sosial.

Penelitian mengenai deteksi cyberbullying dan analisis sentimen telah banyak dilakukan menggunakan berbagai pendekatan *machine learning*. Mahajan dan Sah [15] menunjukkan bahwa kombinasi NLP dan *machine learning* mampu mendeteksi cyberbullying secara efektif pada berbagai platform media sosial. Penelitian lain menunjukkan bahwa klasifikasi berbasis teks dapat digunakan untuk mengidentifikasi pola perilaku cyberbullying melalui komentar pengguna [6]. Selain itu, berbagai studi terbaru juga menunjukkan bahwa teknik *text mining* dan algoritma klasifikasi mampu mengidentifikasi komentar yang berpotensi mengandung ujaran kebencian maupun perilaku agresif pada media sosial [16], [17]. Meskipun demikian, sebagian besar penelitian masih berfokus pada platform Twitter, Facebook, dan Instagram, sementara penelitian yang secara khusus menganalisis komentar TikTok berbahasa Indonesia terkait isu remaja masih relatif terbatas. Selain itu, sebagian besar penelitian terdahulu lebih menitikberatkan pada peningkatan performa model klasifikasi tanpa menghubungkan hasil analisis dengan upaya penguatan literasi digital sebagai strategi preventif terhadap cyberbullying.

Literasi digital merupakan kemampuan individu dalam mengakses, memahami, mengevaluasi, menciptakan, dan menggunakan informasi digital secara kritis, etis, serta bertanggung jawab [18]. Literasi digital tidak hanya berkaitan dengan kemampuan teknis dalam mengoperasikan teknologi, tetapi juga mencakup kemampuan memahami risiko digital, etika komunikasi, keamanan informasi, serta tanggung jawab sosial dalam berinteraksi di ruang digital [19]. Individu yang memiliki tingkat literasi digital yang baik cenderung lebih mampu mengenali perilaku cyberbullying, menghindari keterlibatan dalam tindakan tersebut, serta mengambil langkah yang tepat ketika menjadi korban maupun saksi perundungan daring [20]. Dengan demikian, hasil analisis sentimen yang menunjukkan keberadaan komentar negatif dan indikasi cyberbullying dapat dimanfaatkan sebagai dasar dalam penyusunan program edukasi dan penguatan literasi digital bagi remaja.

Berdasarkan kajian literatur yang telah dilakukan, terdapat beberapa kesenjangan penelitian (*research gap*) yang perlu diperhatikan. Pertama, penelitian mengenai cyberbullying masih didominasi oleh analisis pada platform Twitter, Facebook, dan Instagram, sedangkan kajian terhadap komentar TikTok berbahasa Indonesia terkait isu remaja masih relatif terbatas. Kedua, sebagian besar penelitian berfokus pada aspek teknis peningkatan akurasi model klasifikasi tanpa mengaitkan hasil analisis dengan upaya preventif dalam bentuk penguatan literasi digital. Ketiga, masih sedikit penelitian yang memanfaatkan hasil identifikasi sentimen negatif sebagai dasar penyusunan rekomendasi edukasi digital yang relevan bagi remaja Indonesia. Kesenjangan tersebut menunjukkan perlunya penelitian yang tidak hanya berfokus pada klasifikasi sentimen, tetapi juga pada pemanfaatan hasil klasifikasi sebagai dasar pengembangan strategi pencegahan cyberbullying melalui literasi digital.

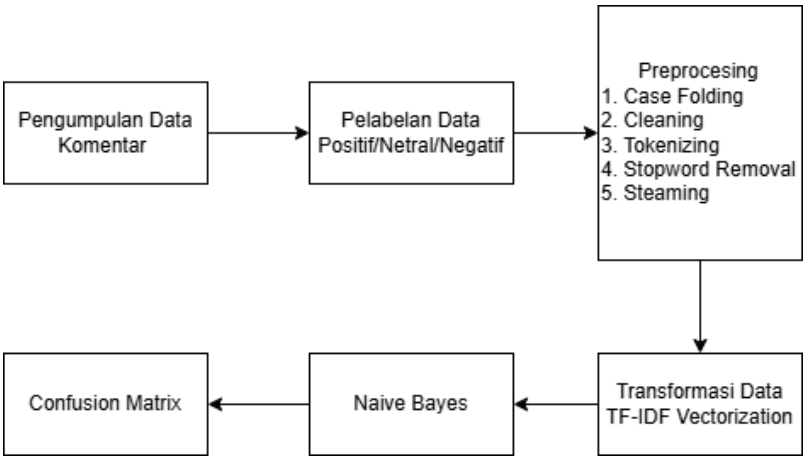
Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis sentimen komentar TikTok terkait isu remaja menggunakan algoritma Multinomial Naïve Bayes, mengidentifikasi dominasi sentimen yang muncul, serta mengkaji implikasi hasil analisis terhadap penguatan literasi digital. Penelitian ini diharapkan dapat memberikan kontribusi teoritis dalam pengembangan kajian analisis sentimen dan deteksi cyberbullying pada media sosial, sekaligus memberikan kontribusi praktis bagi sekolah, orang tua, dan pembuat kebijakan dalam merancang strategi edukasi digital yang lebih efektif guna menciptakan lingkungan digital yang aman, sehat, dan bertanggung jawab bagi remaja.

2 METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain deskriptif-komputasional (computational descriptive research) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan sentimen komentar TikTok terkait isu remaja serta mendeteksi indikasi cyberbullying menggunakan algoritma Multinomial Naïve Bayes. Pendekatan kuantitatif dipilih karena memungkinkan pengukuran objektif terhadap pola sentimen yang muncul dalam data teks dan menghasilkan temuan yang dapat dianalisis secara statistik. Penelitian ini juga mengadopsi pendekatan text mining dan Natural Language Processing (NLP) yang telah banyak digunakan dalam penelitian analisis sentimen untuk mengidentifikasi opini, emosi, dan kecenderungan sikap pengguna berdasarkan data tekstual [21], [22]. Melalui pendekatan tersebut, data komentar dapat diolah menjadi informasi yang lebih terstruktur sehingga memudahkan proses klasifikasi sentimen dan identifikasi indikasi cyberbullying. Hasil klasifikasi selanjutnya digunakan sebagai dasar untuk merumuskan rekomendasi penguatan literasi digital bagi remaja dalam menghadapi fenomena cyberbullying di media sosial..

2.1 Alur Penelitian

Penelitian ini dilakukan melalui tujuh tahapan utama, yaitu pengumpulan data, pelabelan data, preprocessing teks, transformasi data, klasifikasi menggunakan Naïve Bayes Classifier, evaluasi model, serta analisis dan interpretasi hasil. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

2.2 Tahapan Penelitian

2.2.1 Pengumpulan Data

Data penelitian berupa komentar TikTok yang berkaitan dengan isu remaja, seperti kesehatan mental, pendidikan, perundungan, gaya hidup, dan fenomena sosial remaja. Data dikumpulkan menggunakan teknik web scraping dengan memperhatikan kebijakan penggunaan data platform TikTok. Komentar yang berhasil dikumpulkan kemudian disimpan dalam format .csv untuk memudahkan proses pengolahan data.

2.2.2 Pelabelan Data

Komentar TikTok yang telah dikumpulkan diberi label secara manual berdasarkan kategori positif, dan negatif. Kategori positif mencakup komentar yang bersifat mendukung, sedangkan negatif/cyberbullying mencakup komentar yang mengandung penghinaan, ejekan, ujaran kebencian, atau bentuk perundungan lainnya. Alternatif lain, pelabelan dapat menggunakan klasifikasi biner, yaitu cyberbullying dan non-cyberbullying. Untuk memastikan konsistensi hasil pelabelan, dilakukan pengujian inter-rater agreement menggunakan Cohen's Kappa. Dataset yang telah diberi label selanjutnya digunakan sebagai data latih dan data uji dalam proses klasifikasi menggunakan algoritma Naïve Bayes Classifier.

2.2.3 Preprocessing Data

Preprocessing data merupakan tahap awal pengolahan teks yang bertujuan untuk membersihkan dan menyiapkan data komentar TikTok agar dapat diproses oleh algoritma klasifikasi. Tahap ini dilakukan untuk mengurangi noise pada data dan meningkatkan kualitas hasil analisis sentimen. Preprocessing dilakukan menggunakan Python pada Google Colaboratory dengan beberapa tahapan, yaitu case folding, cleaning, tokenizing, stopword removal, dan stemming.

2.2.3.1 Case Folding

Case folding merupakan proses mengubah seluruh karakter huruf pada komentar menjadi huruf kecil (lowercase). Tahap ini bertujuan untuk menyeragamkan format teks sehingga kata yang memiliki makna sama tetapi berbeda penulisan, seperti "TikTok", "TIKTOK", dan "tiktok", akan dianggap sebagai satu kata yang sama oleh sistem, sebagai contoh "TikTok Banyak Digunakan Remaja" menjadi "tiktok banyak digunakan remaja".

2.2.3.2 Cleaning

Cleaning merupakan proses pembersihan data dengan menghapus elemen-elemen yang tidak memiliki kontribusi terhadap analisis sentimen, seperti URL, emoji, tanda baca, angka, hashtag (#), mention (@), dan karakter khusus lainnya. Tahap ini dilakukan untuk mengurangi noise pada data teks sehingga informasi yang dianalisis lebih relevan, sebagai contoh "Remaja sekarang keren banget 😍😍 #fyp" menjadi "remaja sekarang keren banget".

2.2.3.3 Tokenizing

Tokenizing adalah proses memecah kalimat menjadi unit-unit kata (token) yang lebih kecil. Tahap ini memudahkan sistem dalam mengenali dan mengolah setiap kata yang terdapat dalam komentar, sebagai contoh "remaja sekarang keren banget" menjadi ["remaja", "sekarang", "keren", "banget"].

2.2.3.4 Stopword Removal

Stopword removal merupakan proses menghapus kata-kata umum yang sering muncul tetapi tidak memiliki makna penting dalam proses klasifikasi sentimen. Kata-kata seperti "dan", "yang", "di", "ke", "dari", dan "untuk" termasuk dalam kategori stopword, sebagai contoh ["remaja", "yang", "sekarang", "keren", "banget"] menjadi ["remaja", "sekarang", "keren", "banget"].

2.2.3.5 Stemming

Stemming adalah proses mengubah kata berimbuhan menjadi bentuk kata dasarnya. Pada penelitian ini, stemming dilakukan menggunakan pustaka Sastrawi yang umum digunakan untuk pemrosesan teks berbahasa Indonesia. Tujuan stemming adalah mengurangi variasi kata sehingga kata-kata yang memiliki akar kata sama dapat dikenali sebagai satu fitur yang sama.

Tabel 1. Perbedaan Kata Awal dengan Hasil Stemming

Kata Awal	Hasil Stemming
Perundungan	Rundung

Bullyingnya	Bullying
Menghina	Hina
Komentar	Komentar
Digunakan	Guna

2.2.4 Transformasi Data

Setelah melalui tahap preprocessing, data teks yang telah bersih belum dapat langsung diproses oleh algoritma Naïve Bayes Classifier karena masih berbentuk kata atau kalimat. Oleh karena itu, diperlukan proses transformasi data untuk mengubah teks menjadi representasi numerik yang dapat dipahami oleh model klasifikasi. Pada penelitian ini, metode yang digunakan adalah Term Frequency–Inverse Document Frequency (TF-IDF) karena mampu memberikan bobot pada setiap kata berdasarkan tingkat kepentingannya dalam suatu dokumen.

TF-IDF bekerja dengan menghitung frekuensi kemunculan suatu kata (Term Frequency/TF) dalam sebuah dokumen dan mengombinasikannya dengan nilai Inverse Document Frequency (IDF) yang menunjukkan tingkat keunikan kata tersebut dalam keseluruhan dataset. Semakin sering suatu kata muncul pada dokumen tertentu tetapi jarang muncul pada dokumen lain, maka bobot TF-IDF yang dihasilkan akan semakin tinggi. Dengan demikian, kata-kata yang dianggap penting dalam menentukan sentimen akan memiliki pengaruh yang lebih besar dalam proses klasifikasi. Perhitungan nilai TF-IDF dilakukan menggunakan Persamaan (1) dan Persamaan (2).

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

$$IDF(t) = \log \frac{N}{df(t)} \quad (2)$$

Keterangan dari Persamaan (1) dan (2) adalah sebagai berikut:

1) TF(t,d) = frekuensi kemunculan kata t pada dokumen d.
2) IDF(t) = bobot keunikan kata t.
3) N = jumlah seluruh dokumen.
4) df(t) = jumlah dokumen yang mengandung kata t.

Melalui proses pembobotan TF-IDF, fitur teks yang relevan dapat direpresentasikan dengan lebih baik sehingga mampu meningkatkan akurasi model dalam mengidentifikasi sentimen positif dan negatif. Sebagai contoh, hasil preprocessing dari komentar "parah banget hina teman terus" akan diubah menjadi vektor numerik berdasarkan bobot TF-IDF masing-masing kata, seperti parah, banget, hina, teman, dan terus. Hasil transformasi ini berupa matriks fitur (feature matrix) yang kemudian digunakan sebagai input pada proses klasifikasi menggunakan algoritma Naïve Bayes Classifier. Transformasi data dilakukan menggunakan pustaka TfidfVectorizer pada library Scikit-learn di Python melalui platform Google Colaboratory. Dengan metode ini, informasi penting dalam teks dapat direpresentasikan secara lebih efektif sehingga meningkatkan kinerja model klasifikasi sentimen dan deteksi cyberbullying.

2.2.5 Klasifikasi Menggunakan Naive Bayes Classifier

Tahap klasifikasi dilakukan setelah data teks ditransformasikan ke dalam bentuk numerik menggunakan metode TF-IDF. Pada penelitian ini, algoritma yang digunakan adalah Naïve Bayes Classifier (NBC), yaitu metode klasifikasi probabilistik yang bekerja berdasarkan Teorema Bayes. Algoritma ini dipilih karena memiliki tingkat komputasi yang relatif cepat, sederhana, serta mampu memberikan performa yang baik dalam klasifikasi teks dan analisis sentimen pada data media sosial.

Sebelum proses klasifikasi dilakukan, dataset dibagi menjadi dua bagian, yaitu data latih (training data) sebesar 80% dan data uji (testing data) sebesar 20%. Data latih digunakan untuk

membangun model klasifikasi, sedangkan data uji digunakan untuk mengukur kemampuan model dalam mengklasifikasikan data yang belum pernah dipelajari sebelumnya. Naïve Bayes Classifier bekerja dengan menghitung probabilitas suatu komentar termasuk ke dalam kelas tertentu berdasarkan kata-kata yang terkandung di dalamnya. Perhitungan Teorema Bayes dilakukan menggunakan Persamaan (3).

$$P(C \mid X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (3)$$

Keterangan dari Persamaan (3) adalah sebagai berikut:

1) $P(C \mid X)$= probabilitas data Xberada pada kelas C
2) $P(X \mid C)$= probabilitas kemunculan fitur Xpada kelas C
3) $P(C)$= probabilitas awal dari kelas C
4) $P(X)$= probabilitas kemunculan data X

Dalam penelitian ini, kelas yang digunakan terdiri atas positif dan negatif/cyberbullying. Model akan menghitung probabilitas setiap komentar terhadap masing-masing kelas, kemudian menetapkan kelas dengan nilai probabilitas tertinggi sebagai hasil prediksi.

Proses klasifikasi dilakukan menggunakan Multinomial Naïve Bayes, yang merupakan varian Naïve Bayes yang paling sesuai untuk data teks berbasis frekuensi kata atau bobot TF-IDF. Implementasi model dilakukan menggunakan library Scikit-learn pada Python melalui platform Google Colaboratory. Hasil dari tahap ini berupa label prediksi sentimen untuk setiap komentar TikTok yang selanjutnya digunakan pada tahap evaluasi model untuk mengukur tingkat kinerja klasifikasi.

2.2.6 Confusion Matrix

Setelah proses klasifikasi menggunakan Naïve Bayes Classifier selesai dilakukan, tahap selanjutnya adalah mengevaluasi kinerja model menggunakan Confusion Matrix. Confusion Matrix merupakan metode evaluasi yang digunakan untuk membandingkan hasil prediksi model dengan data aktual sehingga dapat diketahui tingkat ketepatan klasifikasi yang dihasilkan. Evaluasi ini penting untuk mengukur kemampuan model dalam mengidentifikasi sentimen komentar TikTok secara akurat.

Confusion Matrix terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). True Positive menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar ke dalam kelas positif, True Negative menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar ke dalam kelas negatif, False Positive menunjukkan data yang sebenarnya negatif tetapi diprediksi positif, sedangkan False Negative menunjukkan data yang sebenarnya positif tetapi diprediksi negatif.

Berdasarkan nilai Confusion Matrix, beberapa metrik evaluasi dihitung untuk mengukur performa model, yaitu Accuracy, Precision, Recall, dan F1-Score. Accuracy digunakan untuk mengetahui persentase prediksi yang benar terhadap seluruh data yang diuji dan dihitung menggunakan persamaan (4)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Keterangan dari persamaan (4) adalah sebagai berikut:

1) TP (True Positive) = jumlah data positif yang berhasil diprediksi positif.
2) TN (True Negative) = jumlah data negatif yang berhasil diprediksi negatif.
3) FP (False Positive) = jumlah data negatif yang diprediksi sebagai positif.
4) FN (False Negative) = jumlah data positif yang diprediksi sebagai negatif.

Precision digunakan untuk mengukur tingkat ketepatan model dalam mengidentifikasi suatu kelas dan dihitung dengan persamaan (5)

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Keterangan dari persamaan (5) adalah sebagai berikut:

1) TP (True Positive) = jumlah data positif yang berhasil diprediksi positif.
2) TN (True Negative) = jumlah data negatif yang berhasil diprediksi negatif.
3) FP (False Positive) = jumlah data negatif yang diprediksi sebagai positif.
4) FN (False Negative) = jumlah data positif yang diprediksi sebagai negatif.

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang termasuk dalam suatu kelas dan dihitung menggunakan persamaan (6)

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Keterangan dari persamaan (6) adalah sebagai berikut:

1) TP (True Positive) = jumlah data positif yang berhasil diprediksi positif.
2) TN (True Negative) = jumlah data negatif yang berhasil diprediksi negatif.
3) FP (False Positive) = jumlah data negatif yang diprediksi sebagai positif.
4) FN (False Negative) = jumlah data positif yang diprediksi sebagai negatif.

Sedangkan F1-Score merupakan rata-rata harmonis antara Precision dan Recall yang dihitung menggunakan persamaan (7)

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Keterangan dari persamaan (7) adalah sebagai berikut:

1) Precision = tingkat ketepatan model dalam mengklasifikasikan suatu kelas.
2) Recall = kemampuan model dalam menemukan seluruh data yang termasuk dalam suatu kelas.

Nilai F1-Score berada pada rentang 0 hingga 1. Semakin mendekati nilai 1, maka semakin baik keseimbangan antara Precision dan Recall yang dimiliki oleh model klasifikasi. Nilai Accuracy, Precision, Recall, dan F1-Score yang tinggi menunjukkan bahwa model Naïve Bayes memiliki kemampuan yang baik dalam mengklasifikasikan sentimen komentar TikTok terkait isu remaja. Seluruh proses evaluasi dilakukan menggunakan library Scikit-learn pada Python, sedangkan hasil Confusion Matrix ditampilkan dalam bentuk tabel atau visualisasi matriks untuk mempermudah interpretasi hasil klasifikasi.

3 HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

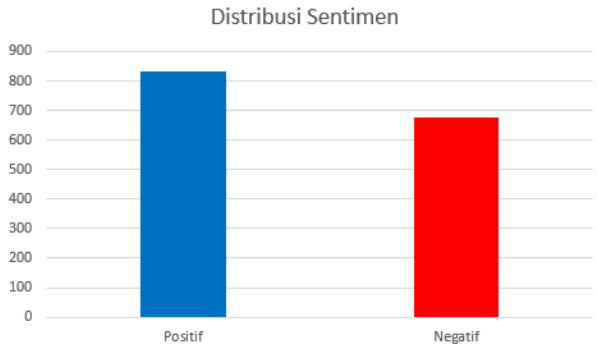
3.1.1 Distribusi Sentimen Komentar TikTok

Analisis distribusi sentimen dilakukan untuk mengetahui kecenderungan opini pengguna TikTok terhadap isu-isu yang berkaitan dengan remaja. Klasifikasi sentimen dilakukan terhadap seluruh komentar yang telah melalui tahapan preprocessing dan pelabelan data. Hasil klasifikasi kemudian dikelompokkan ke dalam dua kategori utama, yaitu sentimen positif dan sentimen negatif. Distribusi jumlah dan persentase masing-masing sentimen dapat dilihat pada Tabel 2.

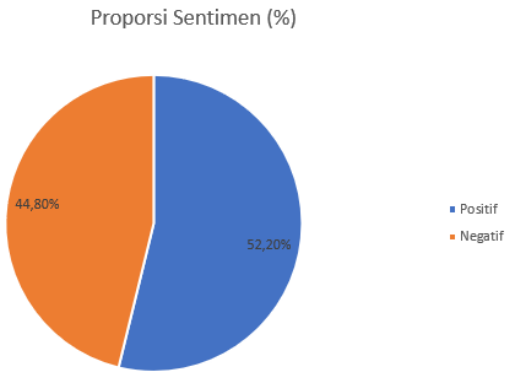
Tabel 2. Distribusi Sentimen Komentar Tiktok

Sentimen	Jumlah	Persentase
Positif	832	55,2%
Negatif	676	44,8%
Total	1.508	100%

Berdasarkan Tabel 2, sentimen positif mendominasi komentar pengguna TikTok dengan jumlah 832 komentar atau sebesar 55,2%, sedangkan sentimen negatif berjumlah 676 komentar atau sebesar 44,8%. Untuk memberikan gambaran yang lebih jelas mengenai distribusi sentimen, hasil klasifikasi divisualisasikan dalam bentuk diagram batang dan diagram lingkaran sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Distribusi Sentimen Komentar TikTok



Gambar 3. Proporsi Sentimen Komentar TikTok

3.1.2 Visualisasi Kata Menggunakan Word Cloud

Visualisasi Word Cloud digunakan untuk menggambarkan kata-kata yang paling sering muncul dalam komentar TikTok terkait isu remaja. Metode ini memberikan representasi visual terhadap frekuensi kemunculan kata pada dataset yang telah melalui tahapan preprocessing. Semakin besar ukuran kata yang ditampilkan, semakin tinggi frekuensi kemunculan kata tersebut dalam komentar yang dianalisis. Visualisasi ini membantu mengidentifikasi topik pembahasan, pola komunikasi pengguna, serta kata-kata yang berpotensi berkaitan dengan sentimen negatif dan cyberbullying.

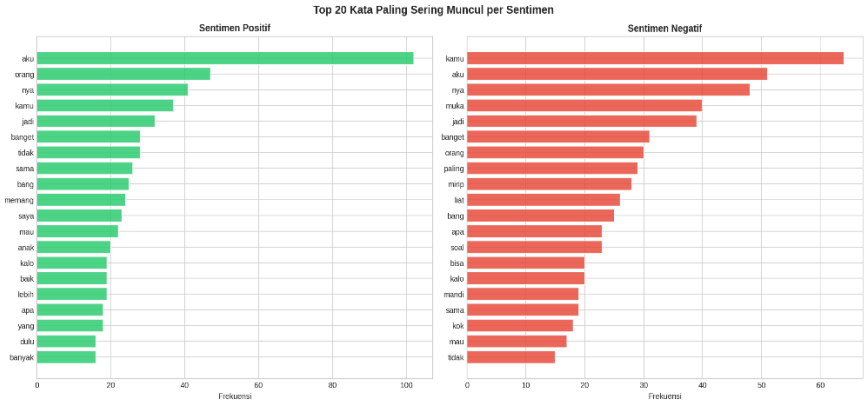
Gambar 4. Word Cloud Komentar TikTok Terkait Isu Remaja

Berdasarkan Gambar 4, terlihat beberapa kata yang memiliki ukuran lebih besar dibandingkan kata lainnya, yang menunjukkan bahwa kata-kata tersebut memiliki frekuensi kemunculan yang lebih tinggi dalam dataset. Dominasi kata-kata tertentu mengindikasikan topik yang paling sering dibahas oleh pengguna TikTok pada isu remaja. Selain itu, kemunculan kata-kata yang bernuansa negatif dapat menjadi indikator adanya perilaku komunikasi yang berpotensi mengarah pada cyberbullying. Hasil visualisasi ini memberikan gambaran awal mengenai karakteristik komentar pengguna sebelum dilakukan analisis lebih lanjut menggunakan algoritma Naïve Bayes Classifier.

Secara keseluruhan, Word Cloud berfungsi sebagai alat eksplorasi data yang membantu memahami pola penggunaan bahasa pada komentar TikTok. Informasi yang diperoleh dari visualisasi ini selanjutnya diperkuat melalui analisis frekuensi kata dan proses klasifikasi sentimen untuk mengidentifikasi kecenderungan sentimen serta potensi cyberbullying pada interaksi pengguna media sosial.

3.1.3 Analisis Frekuensi Kata

Analisis frekuensi kata dilakukan untuk mengidentifikasi kata-kata yang paling sering muncul pada komentar TikTok terkait isu remaja. Analisis ini bertujuan untuk memberikan gambaran mengenai pola penggunaan bahasa oleh pengguna serta mengungkap kata-kata dominan yang berkontribusi terhadap pembentukan sentimen dalam dataset. Setelah melalui tahapan preprocessing yang meliputi case folding, cleaning, tokenizing, stopword removal, dan stemming, setiap kata dihitung frekuensi kemunculannya untuk mengetahui kata yang paling sering digunakan dalam komentar.



Gambar 5. Frekuensi Kata Terbanyak pada Komentar TikTok

Berdasarkan Gambar 5, terlihat bahwa beberapa kata memiliki frekuensi kemunculan yang lebih tinggi dibandingkan kata lainnya. Kata-kata yang sering muncul menunjukkan topik atau isu yang paling banyak dibahas oleh pengguna TikTok dalam konteks remaja. Tingginya frekuensi kemunculan suatu kata mengindikasikan bahwa kata tersebut memiliki peran penting dalam membentuk opini dan respons pengguna terhadap konten yang dianalisis.

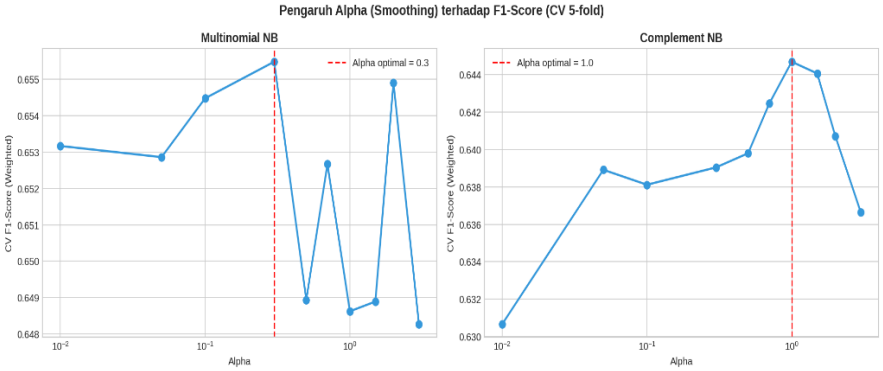
Selain menunjukkan topik yang dominan, analisis frekuensi kata juga dapat digunakan untuk mengidentifikasi potensi kata-kata yang mengandung unsur negatif atau cyberbullying. Kata-kata yang berkonotasi penghinaan, ejekan, atau merendahkan dapat menjadi indikator adanya perilaku perundungan digital dalam interaksi pengguna. Sebaliknya, kemunculan kata-kata yang bersifat mendukung dan memotivasi menunjukkan adanya interaksi positif dalam ruang digital.

Hasil analisis frekuensi kata ini melengkapi visualisasi Word Cloud dengan memberikan informasi yang lebih terukur mengenai jumlah kemunculan setiap kata. Temuan ini selanjutnya menjadi dasar dalam proses klasifikasi sentimen menggunakan algoritma Naïve Bayes Classifier, karena kata-kata yang sering muncul akan memengaruhi pembentukan fitur dan probabilitas pada model klasifikasi. Dengan demikian, analisis frekuensi kata tidak hanya membantu memahami karakteristik komentar pengguna, tetapi juga mendukung interpretasi hasil analisis sentimen dan identifikasi potensi cyberbullying pada platform TikTok.

3.1.4 Optimasi Parameter Alpha pada Naive Bayes

Optimasi parameter alpha dilakukan untuk memperoleh performa terbaik dari algoritma Multinomial Naïve Bayes dalam mengklasifikasikan sentimen komentar TikTok. Parameter alpha berfungsi sebagai smoothing parameter yang digunakan untuk mengatasi permasalahan probabilitas nol (zero probability) pada kata-kata yang jarang muncul dalam data pelatihan. Pemilihan nilai alpha yang tepat dapat meningkatkan kemampuan model dalam melakukan klasifikasi serta mengurangi kesalahan prediksi.

Pada penelitian ini, proses optimasi dilakukan dengan menguji beberapa nilai alpha dan membandingkan performa model berdasarkan metrik evaluasi yang digunakan. Setiap nilai alpha diuji menggunakan dataset yang sama sehingga diperoleh nilai performa yang dapat dibandingkan secara objektif. Hasil pengujian parameter alpha ditampilkan pada Gambar 5.



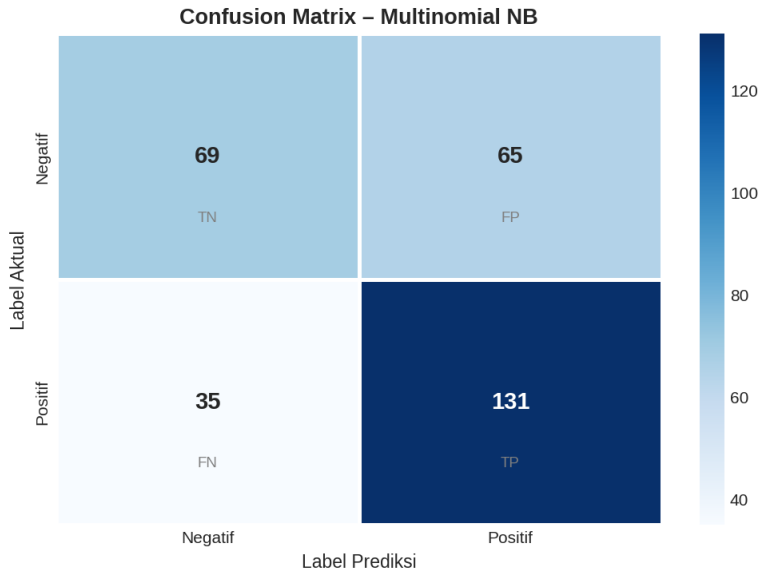
Gambar 6. Hasil Optimasi Parameter Alpha pada Multinomial Naïve Bayes

Berdasarkan Gambar 6, terlihat bahwa perubahan nilai alpha memberikan pengaruh terhadap performa model klasifikasi. Nilai alpha yang terlalu kecil dapat menyebabkan model terlalu sensitif terhadap data pelatihan, sedangkan nilai alpha yang terlalu besar dapat mengurangi kemampuan model dalam membedakan karakteristik antar kelas. Oleh karena itu, nilai alpha terbaik dipilih berdasarkan hasil evaluasi yang menghasilkan performa paling optimal.

Hasil optimasi menunjukkan bahwa penggunaan parameter alpha yang sesuai mampu meningkatkan stabilitas dan akurasi model dalam mengklasifikasikan sentimen komentar TikTok. Temuan ini menunjukkan bahwa proses hyperparameter tuning merupakan tahapan penting dalam penerapan algoritma Naïve Bayes, karena dapat membantu memperoleh model yang lebih efektif dalam mendeteksi sentimen positif maupun negatif. Nilai alpha terbaik yang diperoleh kemudian digunakan pada tahap evaluasi akhir model menggunakan Confusion Matrix, Accuracy, Precision, Recall, dan F1-Score.

3.1.5 Hasil Klasifikasi Menggunakan Naive Bayes Classifier

Setelah proses transformasi data menggunakan TF-IDF selesai dilakukan, tahap berikutnya adalah pembangunan model klasifikasi menggunakan algoritma multinomial naïve bayes. Model yang telah terbentuk kemudian diuji menggunakan data testing untuk mengetahui kemampuan model dalam mengidentifikasi sentimen komentar secara tepat. Hasil klasifikasi divisualisasikan menggunakan Confusion Matrix yang ditunjukkan pada Gambar 7.



Gambar 7. Confusion Matrix Model Multinominal Naïve Bayes

Berdasarkan Gambar 7, model berhasil mengklasifikasikan 69 komentar negatif dan 131 komentar positif dengan benar. Namun, masih terdapat 65 komentar negatif yang diprediksi sebagai positif (*False Positive*) dan 35 komentar positif yang diprediksi sebagai negatif (*False Negative*). Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali pola sentimen pada komentar TikTok, meskipun masih ditemukan beberapa kesalahan klasifikasi.

Berdasarkan perhitungan Confusion Matrix, model menghasilkan accuracy sebesar 69,33%, precision sebesar 68,14%, recall sebesar 83,73%, dan F1-score sebesar 75,14%. Nilai recall yang lebih tinggi dibandingkan precision menunjukkan bahwa model relatif baik dalam mendeteksi komentar yang termasuk ke dalam kelas positif. Secara keseluruhan, hasil evaluasi menunjukkan bahwa algoritma Multinomial Naïve Bayes memiliki performa yang cukup baik untuk digunakan dalam analisis sentimen komentar TikTok terkait isu remaja dan identifikasi potensi cyberbullying.

Tabel 3. Hasil Confusion Matrix

Aktual	Prediksi Negatif	Prediksi Positif
Negatif	69	65
Positif	35	131

3.1.6 Evaluasi Kinerja Model

Evaluasi model dilakukan untuk mengukur tingkat keberhasilan algoritma multinomial naive bayes dalam mengklasifikasikan sentimen komentar TikTok. Pengukuran performa model dilakukan menggunakan beberapa metrik evaluasi, yaitu Accuracy, Precision, Recall, dan F1-Score. Hasil perhitungan masing-masing metrik disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Model Multinomial Naive Bayes

Metrik	Nilai
Accuracy	69,33%
Precision	68,14%
Recall	83,73%
F1-Score	75,14%

Berdasarkan Tabel 4, model menghasilkan nilai accuracy sebesar 69,33%, yang menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan dengan benar. Nilai precision sebesar 68,14% menunjukkan tingkat ketepatan model dalam mengidentifikasi komentar positif, sedangkan recall sebesar 83,73% menunjukkan kemampuan model dalam menemukan komentar positif yang sebenarnya. Sementara itu, F1-Score sebesar 75,14% menunjukkan keseimbangan yang cukup baik antara precision dan recall sehingga model dapat digunakan untuk klasifikasi sentimen pada komentar TikTok terkait isu remaja.

3.2 Pembahasan

Hasil analisis sentimen menunjukkan bahwa dari 1.508 komentar TikTok yang dianalisis, sebanyak 832 komentar (55,2%) tergolong sentimen positif dan 676 komentar (44,8%) tergolong sentimen negatif. Temuan ini menunjukkan bahwa interaksi pengguna TikTok terkait isu remaja cenderung didominasi oleh respons yang positif. Namun, persentase sentimen negatif yang masih cukup tinggi mengindikasikan bahwa ruang digital tetap memiliki potensi munculnya komentar yang mengandung unsur penghinaan, ejekan, maupun perilaku cyberbullying. Kondisi tersebut menunjukkan bahwa media sosial tidak hanya menjadi sarana komunikasi dan berbagi informasi, tetapi juga dapat menjadi ruang terjadinya interaksi negatif yang berpotensi memengaruhi kondisi psikologis remaja.

Hasil visualisasi Word Cloud dan analisis frekuensi kata menunjukkan adanya sejumlah kata yang sering digunakan oleh pengguna dalam memberikan tanggapan terhadap konten yang berkaitan dengan isu remaja. Dominasi kata-kata tertentu mengindikasikan topik yang menjadi perhatian utama pengguna TikTok. Selain itu, kemunculan beberapa kata yang berkonotasi negatif menunjukkan bahwa masih terdapat pola komunikasi yang berpotensi mengarah pada cyberbullying. Temuan ini sejalan dengan konsep online disinhibition effect yang menyatakan bahwa pengguna media sosial cenderung lebih bebas dalam mengekspresikan pendapatnya dibandingkan komunikasi tatap muka, sehingga meningkatkan kemungkinan munculnya komentar yang bersifat agresif atau merugikan pihak lain.

Dari sisi performa model, algoritma Multinomial Naïve Bayes menghasilkan nilai accuracy sebesar 69,33%, precision sebesar 68,14%, recall sebesar 83,73%, dan F1-score sebesar 75,14%. Nilai recall yang lebih tinggi dibandingkan precision menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mendeteksi komentar pada kategori target, meskipun masih terdapat sejumlah kesalahan klasifikasi. Hasil ini menunjukkan bahwa algoritma Naïve Bayes mampu mengidentifikasi pola sentimen pada komentar TikTok dengan performa yang cukup baik, terutama karena karakteristik data berupa teks yang sesuai dengan pendekatan probabilistik yang digunakan oleh algoritma tersebut.

Temuan penelitian ini sejalan dengan penelitian Al-Garadi et al. (2022) yang menyatakan bahwa algoritma Naïve Bayes masih menjadi salah satu metode klasifikasi teks yang efektif untuk analisis sentimen pada media sosial karena memiliki proses komputasi yang sederhana dan efisien. Hasil penelitian ini juga mendukung penelitian Fitro et al. (2024) yang menemukan bahwa metode klasifikasi berbasis machine learning mampu digunakan untuk mendeteksi komentar negatif dan indikasi cyberbullying pada platform media sosial. Meskipun demikian, nilai akurasi yang diperoleh masih berada pada kategori sedang jika dibandingkan dengan beberapa penelitian yang menggunakan model berbasis deep learning seperti LSTM atau Transformer yang umumnya menghasilkan tingkat akurasi lebih tinggi karena mampu memahami konteks kalimat secara lebih mendalam.

Dari perspektif literasi digital, temuan penelitian ini menunjukkan bahwa masih tingginya jumlah komentar negatif pada platform TikTok menjadi indikator perlunya peningkatan kesadaran pengguna, khususnya remaja, mengenai etika berkomunikasi di ruang digital. Literasi digital tidak hanya berkaitan dengan kemampuan mengakses dan menggunakan teknologi, tetapi juga mencakup kemampuan memahami dampak sosial dari setiap aktivitas komunikasi yang dilakukan secara daring. Oleh karena itu, hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam merancang program edukasi digital yang berfokus pada pencegahan cyberbullying, penguatan etika bermedia sosial, serta peningkatan kemampuan berpikir kritis dalam berinteraksi di lingkungan digital.

Penelitian ini memberikan kontribusi praktis bagi sekolah, orang tua, dan pembuat kebijakan dalam memahami pola sentimen dan potensi cyberbullying pada media sosial yang banyak digunakan oleh remaja. Informasi yang diperoleh dari hasil analisis sentimen dapat dimanfaatkan untuk menyusun strategi edukasi yang lebih tepat sasaran, termasuk kampanye anti-cyberbullying dan pengembangan sistem pemantauan komentar berbasis kecerdasan buatan. Dengan demikian, hasil penelitian tidak hanya berkontribusi pada pengembangan kajian analisis sentimen, tetapi juga mendukung upaya menciptakan lingkungan digital yang lebih aman dan sehat bagi remaja.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Data penelitian hanya diperoleh dari platform TikTok sehingga hasil penelitian belum dapat menggambarkan kondisi cyberbullying pada seluruh media sosial. Selain itu, analisis yang dilakukan masih berfokus pada data teks dan belum mempertimbangkan unsur multimodal seperti gambar, video, emoji, maupun konteks sarkasme yang sering ditemukan dalam komunikasi digital. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, mengintegrasikan

analisis multimodal, serta membandingkan performa Naïve Bayes dengan algoritma lain seperti Support Vector Machine (SVM), Random Forest, maupun IndoBERT untuk memperoleh hasil klasifikasi yang lebih akurat dan komprehensif.

4 KESIMPULAN

Penelitian ini bertujuan untuk menganalisis sentimen komentar TikTok terkait isu remaja menggunakan algoritma Multinomial Naïve Bayes sebagai dasar penguatan literasi digital. Hasil penelitian menunjukkan bahwa dari 1.508 komentar yang dianalisis, sebanyak 832 komentar (55,2%) tergolong sentimen positif dan 676 komentar (44,8%) tergolong sentimen negatif. Temuan ini menunjukkan bahwa interaksi pengguna TikTok terkait isu remaja cenderung didominasi oleh sentimen positif, namun masih ditemukan proporsi komentar negatif yang cukup tinggi dan berpotensi mengandung unsur cyberbullying. Dengan demikian, penelitian ini berhasil mengidentifikasi pola sentimen pengguna TikTok serta menunjukkan adanya potensi perilaku perundungan digital dalam interaksi di media sosial.

Berdasarkan hasil evaluasi model, algoritma Multinomial Naïve Bayes menghasilkan nilai accuracy sebesar 69,33%, precision sebesar 68,14%, recall sebesar 83,73%, dan F1-score sebesar 75,14%. Hasil tersebut menunjukkan bahwa algoritma Naïve Bayes memiliki kemampuan yang cukup baik dalam mengklasifikasikan sentimen komentar TikTok. Dengan demikian, hipotesis penelitian yang menyatakan bahwa algoritma Naïve Bayes dapat digunakan untuk mengidentifikasi sentimen komentar dan mendeteksi potensi cyberbullying pada platform TikTok dapat diterima.

Secara praktis, hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam pengembangan program literasi digital bagi remaja, khususnya yang berkaitan dengan etika berkomunikasi dan pencegahan cyberbullying di media sosial. Selain itu, temuan penelitian dapat menjadi masukan bagi sekolah, orang tua, maupun pembuat kebijakan dalam merancang strategi edukasi dan kampanye digital yang mendorong terciptanya lingkungan media sosial yang lebih aman, sehat, dan bertanggung jawab.

Penelitian ini memiliki keterbatasan pada penggunaan data yang hanya berasal dari platform TikTok serta analisis yang masih berfokus pada data teks tanpa mempertimbangkan unsur visual, emoji, maupun konteks bahasa seperti sarkasme. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih luas, mencakup berbagai platform media sosial, serta membandingkan performa Naïve Bayes dengan algoritma lain yang lebih mutakhir seperti Support Vector Machine (SVM), Long Short-Term Memory (LSTM), atau IndoBERT untuk meningkatkan akurasi dan pemahaman terhadap fenomena cyberbullying di ruang digital.

REFERENSI

[1] W. A. Social and Meltwater, "Digital 2024: Global Overview Report," 2024.

[2] APJII, "Laporan Survei Internet Indonesia 2024," 2024.

[3] M. Anderson and J. Jiang, "Teens, Social Media and Technology 2023," Pew Research Center, 2023.

[4] D. B. V Kaye, X. Chen, and J. Zeng, "The Co-Evolution of TikTok and Social Media," *Media International Australia*, 2024.

[5] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Cyberbullying and Hate Speech in Online Social Networks," *ACM Comput. Surv.*.

[6] D. Wulansari and M. Pratama, "Analysis of Cyberbullying Behavior on Social Networking Platforms Using Machine Learning," *Heliyon*, 2023.

[7] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth," *Psychol. Bull.*, vol. 140, no. 4, pp. 1073–1137, 2014, doi: 10.1037/a0035618.

[8] UNICEF, "The State of the World's Children 2024: Digital Opportunities and Risks for Adolescents," 2024.

[9] A. Bozyigit, "Sentiment Analysis on Social Media: A Brief Survey on Current Approaches," *Soc. Netw. Anal. Min.*, 2021.

[10] K. Rosa and M. Santos, "Social Media Sentiment Analysis: Methods, Challenges, and Applications," *Applied Sciences*, 2021.

[11] B. Mathew, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *AAAI Conference on Artificial Intelligence*, 2021.

[12] Y. Zhang, R. Jin, and Z. Zhou, "Understanding the Effectiveness of Naive Bayes for Text Classification," *Expert Syst. Appl.*, 2021.

[13] T. Haryanto and A. Nugraha, "Indonesian Text Classification Using Multinomial Naive Bayes and TF-IDF Weighting," *Procedia Comput. Sci.*, 2023.

[14] N. Kurniasih, A. Wibowo, and R. Setiawan, "Sentiment Analysis of Indonesian Social Media Comments Using TF-IDF and Naive Bayes," *J. Big Data*, 2022.

[15] P. Mahajan and S. Sah, "Leveraging NLP and Machine Learning to Detect Cyberbullying on Digital Platforms," Springer, 2023.

[16] L. Cheng, J. Li, and Y. Silva, "Neural Approaches to Cyberbullying Detection: A Survey," *ACM Transactions on the Web*, 2022.

[17] A. Kumar and N. Sachdeva, "Machine Learning Techniques for Cyberbullying Detection: A Systematic Review," *Expert Syst. Appl.*, 2023.

[18] UNESCO, "Global Education Monitoring Report 2023: Technology in Education," UNESCO Publishing, 2023.

[19] M. Spante and S. Hashemi, "Digital Competence and Digital Literacy: A Systematic Review," *Educ. Inf. Technol. (Dordr).*, 2021.

[20] P. Sharma and A. Verma, "Digital Literacy and Online Safety Among Adolescents in Social Media Environments," *Comput. Educ.*, 2024.

[21] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2020.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Stanford, CA, USA: Stanford University, 2023.