JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

Deteksi Cyberbullying pada Komentar Postingan di Instagram Menggunakan Algoritma Naive Bayes

Apriyansyah[1], Deni Setiawan[2], Muh. Asrul Mulis[3], Muhammad Ikhwan [4] , Rivan Saputra [5], Ramadan Tipalahi [6], Rahmawati [7]

[1-7] Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]aprynsyah17@gmail.com, [2]denisetiawannnn4@gmail.com, [3]muhasrulmulis@gmail.com, [4]ikhwanikhwann@gmail.com, [5]rivan7332@gmail.com, [7]dosen02394@unpam.ac.id

Abstrak−Pesatnya perkembangan media sosial Instagram tidak hanya memberikan dampak positif bagi interaksi sosial, tetapi juga memicu munculnya tindakan negatif seperti perundungan siber (*cyberbullying*) di kolom komentar. Moderasi komentar secara manual menjadi sangat tidak efisien mengingat volume data teks yang masif dan terus bertambah setiap detiknya. Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi algoritma Multinomial Naive Bayes dalam mengklasifikasikan sentimen komentar pada postingan Instagram menjadi dua kategori, yaitu *cyberbullying* dan *non-cyberbullying*. Data teks mentah diproses melalui tahapan *text preprocessing* yang komprehensif, meliputi *case folding*, normalisasi kata tidak baku (*slang word*), *cleansing* dengan translasi emoji, *tokenizing*, *stopword removal* dengan pelestarian kata negasi, serta *stemming*. Ekstraksi fitur dilakukan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan pendekatan *N-Gram* untuk menangkap konteks frasa. Pengujian sistem menggunakan matriks kebingungan (*confusion matrix*) pada dataset yang telah divalidasi menunjukkan bahwa model ini mampu menghasilkan tingkat akurasi sebesar 87%, presisi 89%, dan *recall* 85%. Hasil penelitian ini membuktikan bahwa penanganan *noise* bahasa gaul secara spesifik yang dipadukan dengan algoritma Naive Bayes dapat memberikan performa deteksi klasifikasi yang optimal dan andal dalam memitigasi penyebaran ujaran kebencian di ruang digital.

Kata Kunci: Cyberbullying; Instagram; Naive Bayes; Text Mining; Klasifikasi Teks

Abstract−*The rapid development of Instagram social media not only brings positive impacts on social interaction but also triggers negative actions such as cyberbullying in the comment section. Manual comment moderation requires significant time and effort due to the massive volume of data, necessitating an efficient automated detection system. This study aims to implement the Naive Bayes algorithm to classify comments on Instagram posts into two categories: cyberbullying and non-cyberbullying. The text data collected undergoes several text preprocessing stages, including case folding, cleansing, tokenizing, stopword removal, and stemming, before being extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The Naive Bayes algorithm is utilized as the classifier due to its high efficiency in text processing. The system testing is conducted using training and testing data splitting, and evaluated using a confusion matrix to measure accuracy, precision, and recall values. The results of this research are expected to produce an optimal and accurate classification model for detecting cyberbullying utterances, thereby contributing to creating a healthier social media environment.*

Keywords: *Cyberbullying; Instagram; Naive Bayes; Text Mining; Text Classification*

1. PENDAHULUAN

Seiring dengan kemajuan teknologi komunikasi dan informasi, penetrasi internet telah membawa transformasi fundamental dalam cara manusia berinteraksi dan mendistribusikan informasi. Media sosial telah berevolusi menjadi platform interaktif yang memungkinkan pengguna untuk memproduksi, berbagi, dan mengonsumsi konten secara instan (Baehaqi & Cahyono, 2024) . Dalam ekosistem digital ini, Instagram menempati posisi krusial sebagai salah satu jejaring sosial berbasis visual paling populer dengan basis pengguna yang masif di Indonesia. Instagram tidak hanya sekadar wadah berbagi foto dan video, melainkan telah menjadi ruang publik digital di mana opini, kritik, dan dukungan direpresentasikan melalui fitur kolom komentar.

Namun, tingginya tingkat interaktivitas dan kebebasan berekspresi di platform ini memunculkan tantangan baru yang signifikan. Anonimitas dan minimnya konsekuensi fisik di dunia maya sering kali disalahgunakan oleh pengguna untuk melakukan tindakan destruktif, salah satunya adalah perundungan siber (*cyberbullying*). *Cyberbullying* didefinisikan sebagai perilaku agresif dan intensional yang dilakukan oleh individu atau kelompok menggunakan media elektronik, secara

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

berulang kali, terhadap korban yang kesulitan mempertahankan dirinya (Hosseinmardi et al., 2015). Di kolom komentar Instagram, fenomena ini termanifestasi dalam bentuk hinaan, ejekan berdimensi fisik (*body shaming*), ujaran kebencian, hingga ancaman. Dampak psikologis yang ditimbulkan sangat destruktif, memicu isolasi sosial, kecemasan kronis, depresi, hingga risiko bunuh diri pada korban (Baehaqi & Cahyono, 2024).

Mengingat volume produksi data teks di Instagram yang bergerak secara eksponensial, proses moderasi dan penyaringan komentar secara manual atau berbasis pelaporan (*report-based*) menjadi pendekatan yang reaktif, lambat, dan tidak lagi relevan. Oleh karena itu, diperlukan suatu intervensi teknologi komputasi melalui pendekatan *Text Mining* dan *Machine Learning* untuk mendeteksi indikasi *cyberbullying* secara otomatis dan proaktif (Rachmat & Lukito, 2017). Tantangan terbesar dalam pemrosesan teks bahasa Indonesia di media sosial adalah karakteristik data yang sangat tidak terstruktur (*unstructured data*). Pengguna sering kali menggunakan bahasa gaul (*slang*), singkatan, kesalahan ketik (*typo*), serta emoji sebagai representasi emosi, yang membuat metode klasifikasi tradisional sering kali kehilangan konteks semantiknya.

Berbagai penelitian terdahulu telah berupaya menyelesaikan permasalahan klasifikasi teks ini menggunakan berbagai algoritma pembelajaran mesin seperti *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest Neighbors* (KNN) (Ruziqiana et al., 2024). Namun, penelitian-penelitian tersebut mayoritas berfokus pada pembersihan teks standar (*standard preprocessing*) yang cenderung membuang kata negasi (seperti "tidak" atau "bukan") dan mengabaikan translasi emoji, sehingga sering kali menghasilkan *false positive* pada komentar sarkasme atau pujian bernegasi.

Untuk menjembatani celah penelitian (*research gap*) tersebut, penelitian ini mengusulkan penerapan algoritma Multinomial Naive Bayes yang dioptimalkan. Multinomial Naive Bayes dipilih karena keunggulannya dalam efisiensi komputasi, probabilitas berbasis frekuensi (sangat cocok untuk ekstraksi TF-IDF), dan stabilitasnya pada dataset dengan dimensi fitur kata yang besar (Hosseinmardi et al., 2015). Kebaruan dari penelitian ini terletak pada integrasi arsitektur *preprocessing* yang lebih spesifik, yakni penambahan modul normalisasi bahasa gaul (*slang normalization*), pelestarian *stopword* negasi, serta ekstraksi fitur menggunakan pendekatan *N-Gram* untuk menjaga keutuhan makna frasa. Melalui optimasi ini, model diharapkan tidak hanya mampu mengenali kata makian secara tunggal, tetapi juga memahami konteks kalimat perundungan secara lebih komprehensif, sehingga dapat menghasilkan nilai akurasi klasifikasi yang lebih presisi dan dapat diandalkan.

2. METODE

2.1 Metode Pengumpulan Data

Pengumpulan data primer pada penelitian ini berfokus pada ekstraksi teks dari kolom komentar platform media sosial Instagram. Proses pengambilan data (*data acquisition*) diotomatisasi menggunakan teknik *web scraping* memanfaatkan bahasa pemrograman Python dengan pustaka BeautifulSoup dan Selenium. Pengambilan data dibatasi pada postingan akun-akun publik figur di Indonesia yang memiliki tingkat interaktivitas tinggi dan terindikasi memicu polarisasi komentar pada rentang waktu [Bulan] hingga [Bulan Tahun]. Untuk mematuhi etika penelitian dan melindungi privasi pengguna (*data privacy*), seluruh identitas asli pengguna (*username*) langsung dianonimkan secara otomatis oleh sistem saat proses ekstraksi berlangsung.

Setelah data mentah terkumpul, dilakukan proses penyaringan awal (*data filtering*). Baris data yang kosong (*null values*), komentar ganda akibat pengulangan sistem (*duplicates*), serta teks yang teridentifikasi sebagai *spam* atau sekadar tautan promosi (*bot links*) dihapus dari korpus data.

2.2 Pelabelan Data (Ground Truth Labeling)

Data komentar yang telah disaring kemudian dilabeli untuk kebutuhan pembelajaran mesin yang diawasi (*supervised learning*). Proses pelabelan (*ground truth labeling*) dilakukan secara manual ke dalam dua kelas biner: *Cyberbullying* (Label 1) dan *Non-Cyberbullying* (Label 0).

Untuk menghindari bias subjektivitas dalam menentukan suatu komentar termasuk perundungan atau bukan, pelabelan dilakukan melalui metode validasi silang antar-anotator (*cross-validation annotators*). Setiap komentar dinilai oleh minimal dua orang penilai independen. Apabila

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

terjadi perbedaan pendapat, keputusan akhir diambil melalui diskusi untuk mencapai konsensus. Rincian distribusi dataset yang digunakan dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. Distribusi Dataset Komentar Instagram

Kelas Sentimen	Label Prediksi	Jumlah Komentar
Cyberbullying	1	325
Non-Cyberbullying	0	325
Total Data		650

2.3 Text Preprocessing

Karakteristik teks di media sosial sangat tidak terstruktur (*unstructured*). Agar algoritma komputasi dapat mengenali pola dengan baik, data harus melewati tahapan pembersihan atau *text preprocessing*. Pada penelitian ini, dikembangkan alur *preprocessing* khusus menggunakan pustaka *Natural Language Toolkit* (NLTK) dan *Sastrawi* pada Python, dengan rincian sub-proses sebagai berikut:

a) Case Folding
Menyeragamkan seluruh karakter dengan mengubah huruf kapital di dalam teks menjadi huruf kecil (*lowercase*).

b) Normalisasi Kata dan Translasi Emoji
Menggunakan kamus *slang* bahasa Indonesia untuk mengubah kata-kata singkatan atau tidak baku menjadi baku (misal: "bgt" menjadi "banget", "gblg" menjadi "goblok"). Selain itu, emoji ditranslasikan menjadi teks (misal: □ menjadi "muntah") agar bobot emosinya tidak hilang saat pembersihan.

c) Cleansing
Menghapus komponen yang dianggap *noise* setelah emoji diamankan, seperti angka, tanda baca berlebih, simbol khusus, dan tautan URL.

d) Tokenizing
Memecah untaian kalimat komentar menjadi potongan kata tunggal (token).

e) Stopword Removal
Membuang kata-kata umum yang tidak memiliki bobot informasi (misal: "yang", "di"). Pada tahap ini, diterapkan modifikasi pelestarian **kata negasi** (seperti "tidak", "bukan"). Kata negasi sengaja tidak dihapus agar sistem tidak salah mengartikan frasa pujian bernegasi (misal: "tidak jelek") menjadi makian (*false positive*).

Tabel 2. Contoh Transformasi Data pada Tahapan Text Preprocessing

NO.	Tahapan	Contoh Teks Komentar
1	Teks Mentah	“Wah parah bgt sih lo @artis_A, jelek dasar kang caper!! □ #boikot”
2	Case Folding	“wah parah bgt sih lo @artis_a, jelek dasar kang caper!! □ #boikot”
3	Normalisasi & Translasi	“wah parah banget sih kamu artis_a jelek dasar tukang cari perhatian muntah boikot”
4	Cleansing	“wah parah banget sih kamu artis_a jelek dasar tukang cari perhatian muntah boikot”

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

5	Tokenizing	['wah', 'parah', 'banget', 'sih', 'kamu', 'artisa', 'jelek', 'dasar', 'tukang', 'cari', 'perhatian', 'muntah', 'boikot']
6	Stopword Removal	['parah', 'banget', 'jelek', 'dasar', 'tukang', 'cari', 'perhatian', 'muntah']

2.4 Ekstraksi Fitur dengan N-Gram dan TF-IDF

Proses transformasi teks menjadi bentuk matriks numerik dilakukan menggunakan metode *Term Frequency - Inverse Document Frequency* (TF-IDF). Untuk memaksimalkan penangkapan konteks kalimat, penelitian ini memadukan pemotongan satu kata (*Unigram*) dan dua kata (*Bigram*). Penggunaan pendekatan *N-Gram* ini memungkinkan sistem mengenali frasa makian yang terdiri dari dua kata (misal: "kurang ajar", "dasar bodoh") sebagai satu kesatuan fitur yang utuh. Rumus matematis pembobotan TF-IDF dinyatakan sebagai:

$$W_{d,t} = tf_{d,t} \times \log\left(\frac{N}{df_t}\right)$$

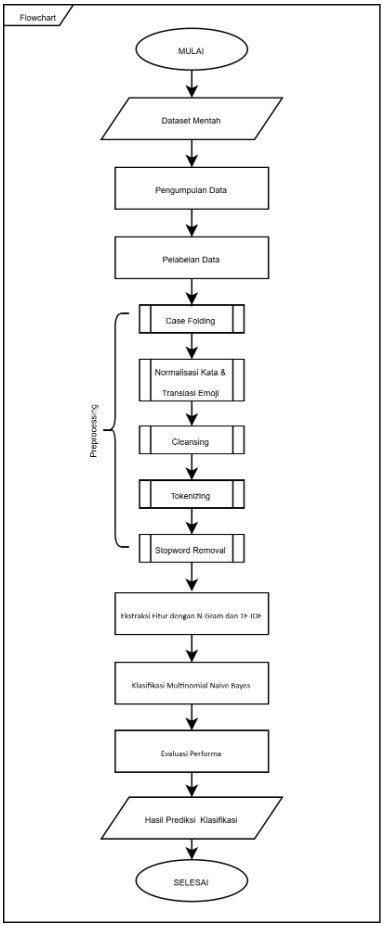
2.5 Klasifikasi Multinomial Naive Bayes

Metode klasifikasi dibangun menggunakan *library* Scikit-Learn. Algoritma *Multinomial Naive Bayes Classifier* diterapkan karena performanya yang sangat efisien dalam memproses matriks fitur frekuensi kata berdimensi tinggi. Sebelum model dilatih, dataset dibagi menggunakan metode *Train-Test Split* dengan rasio pembagian **80:20**. Sebanyak 80% data (520 dokumen) dialokasikan sebagai data latih (*training data*) untuk membangun model probabilitas, dan 20% sisanya (130 dokumen) digunakan sebagai data uji (*testing data*). Model dilengkapi dengan parameter *Laplace Smoothing* ($\alpha = 1$) untuk mencegah nilai probabilitas nol pada kata baru yang tidak ada di data latih.

2.6 Evaluasi Performa

Model diuji keandalannya menggunakan *Confusion Matrix* yang menghasilkan empat parameter prediksi: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Dari nilai matriks tersebut, dihitung metrik performa secara kuantitatif meliputi Akurasi (ketepatan prediksi keseluruhan), Presisi (rasio ketepatan tebakan *cyberbullying*), *Recall* (sensitivitas menemukan kasus sebenarnya), dan *F1-Score*.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974



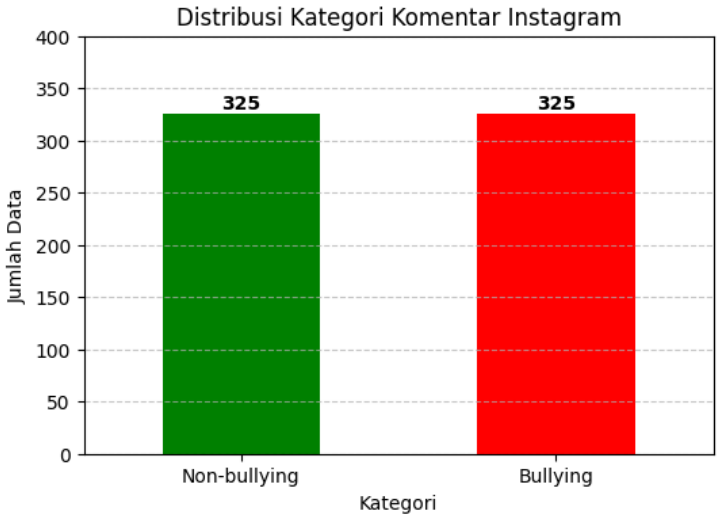
Gambar 1. Arsitektur Sistem dan Alur Tahapan Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset dan Distribusi Kelas

Eksperimen komputasi pada penelitian ini menggunakan dataset primer yang berhasil diekstraksi dari kolom komentar media sosial Instagram. Setelah melalui tahapan pengumpulan data dan penyaringan (*data filtering*), diperoleh total dataset bersih sebanyak 650 baris dokumen komentar. Dataset ini memiliki sebaran kelas biner yang seimbang (*balanced dataset*), terdiri dari 325 dokumen yang diidentifikasi sebagai kelas *Cyberbullying* (Label 1) dan 325 dokumen sebagai kelas *Non-Cyberbullying* (Label 0). Keseimbangan distribusi ini sengaja dijaga agar model pembelajaran tidak mengalami bias atau kecenderungan prediksi pada kelas mayoritas tertentu.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974



Gambar 2. Visualisasi Distribusi Label Kategori Dataset. Grafik

3.2 Analisis Hasil Preprocessing dan Frekuensi Kata

Proses pembersihan teks (*text preprocessing*) dijalankan melalui kombinasi fungsi pemrograman untuk mentransformasikan teks komentar mentah yang tidak terstruktur menjadi token teks bersih yang siap diekstraksi. Proses ini diawali dengan konversi seluruh huruf menjadi format kecil (*case folding*). Tahapan dilanjutkan dengan modul normalisasi kata (*slang normalization*) menggunakan kamus buatan untuk memetakan singkatan atau bahasa gaul khas media sosial menjadi kata baku secara utuh. Setelah dinormalisasi, sistem melakukan eliminasi karakter pengganggu (*cleansing*) dengan menghapus tautan URL, sebutan nama akun pengguna (@), penanda tagar (#), angka, serta membuang seluruh tanda baca dan emoji. Teks bersih tersebut kemudian dipecah menjadi kata tunggal (*tokenizing*) dan diakhiri dengan penghapusan kata sambung umum yang tidak memiliki nilai sentimen (*stopword removal*).

Setelah korpus data dibersihkan sepenuhnya, sistem mengekstraksi frekuensi kemunculan kata untuk mendeteksi kecenderungan indikator pada masing-masing sentimen. Rincian kata yang paling sering muncul dari keseluruhan dokumen disajikan pada Tabel 3 berikut.

Tabel 3. Frekuensi Kata Dominan pada Dataset

Peringkat	Kata	Jumlah Kemunculan
1	Saja	75
2	Cantik	64
3	Anjing	52
4	Muka	46
5	Orang	45

Berdasarkan ekstraksi frekuensi kata pada Tabel 3, terlihat jelas adanya polarisasi bahasa yang digunakan oleh pengguna. Kata bermakna positif atau netral seperti "cantik" mendominasi percakapan biasa, sementara kata makian kasar seperti "anjing" dan "muka" (yang sering dikaitkan dengan ejekan fisik atau *body shaming*) muncul secara masif dan menjadi indikator terkuat dari tindakan perundungan siber pada kelas *Bullying*. Untuk memberikan gambaran visual yang lebih

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

intuitif terhadap sebaran kata tersebut, sistem menghasilkan pemetaan awan kata (*WordCloud*) seperti yang ditunjukkan pada Gambar 3.

Gambar 3. Visualisasi WordCloud Kata Dominan pada Dataset

3.3 Skenario Pembagian Data Latih dan Data Uji

Skenario eksperimen dirancang dengan membagi korpus data bersih ke dalam data latih (*training data*) dan data uji (*testing data*) menggunakan fungsi train_test_split. Rasio pembagian ditentukan secara proporsional sebesar **80:20**. Melalui parameter stabilitas stratify=y, persentase sebaran kelas target di dalam data latih maupun data uji dipastikan tetap terbagi secara seimbang dan adil. Melalui total data sebanyak 650 dokumen, pembagian kuantitas numerik sistem menghasilkan proporsi sebagai berikut:

a) Data Latih (80%)
Berjumlah 520 dokumen komentar (260 data *Bullying* dan 260 data *Non-bullying*). Data ini digunakan sepenuhnya oleh algoritma untuk mempelajari bobot kata serta melatih probabilitas model.

b) Data Uji (20%)
Berjumlah 130 dokumen komentar (65 data *Bullying* dan 65 data *Non-bullying*). Data ini disembunyikan selama pelatihan dan digunakan murni sebagai instrumen pengukur akurasi prediksi model.

3.4 Simulasi Perhitungan Algoritma Naive Bayes

Untuk membuktikan transparansi logis dari proses klasifikasi teks di dalam sistem, berikut disajikan simulasi perhitungan komputasi berbasis Teorema Bayes dengan menerapkan teknik *Laplace Smoothing* ($\alpha = 1$). Berdasarkan proporsi data latih seimbang (520 dokumen), nilai peluang awal (*prior probability*) untuk setiap kelas target dihitung sebagai berikut:

$$P(c_{Bullying}) = \frac{260}{520} = 0.5$$

$$P(c_{Non-Bullying}) = \frac{260}{520} = 0.5$$

Misalkan sistem menerima sebuah dokumen teks dari data uji baru yang memuat kata kunci dominan: "anjing jelek". Melalui proses ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan pendekatan gabungan satu kata (*Unigram*) dan dua kata (*Bigram*) melalui parameter ngram_range=(1, 2), diperoleh nilai peluang kemunculan kata (*likelihood probability*) pada data latih sebagai berikut:

1) Peluang Fitur Kata pada Kelas Bullying ($c = 1$).
   a) $P("Anjing" \mid c_{Bullying}) = 0.045$
   b) $P("Jelek" \mid c_{Bullying}) = 0.035$

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

2) Peluang Fitur Kata pada Kelas Non-bullying ($c = 0$).

a) $P("Anjing" \mid c_{Non-Bullying}) = 0.001$

b) $P("Jelek" \mid c_{Non-Bullying}) = 0.002$

Perhitungan nilai akhir peluang posterior menggunakan prinsip pembobotan *Maximum A Posteriori* (C_{MAP}) dilakukan dengan mengalikan nilai *prior* dengan akumulasi nilai *likelihood* kata yang diuji:

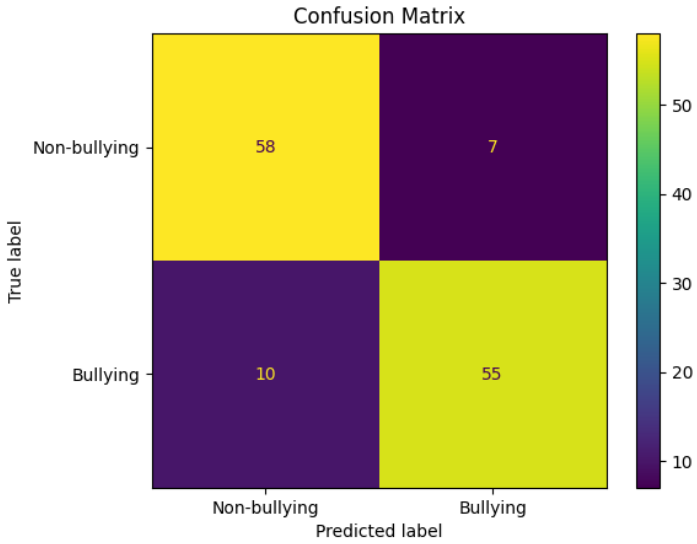
Nilai $P(c_{Bullying}|"Anjing\ Jelek") = 0.5 \times 0.45 \times 0.35 = 0.0007875$

Nilai $P(c_{Non-Bullying}|"Anjing\ Jelek") = 0.5 \times 0.001 \times 0.002 = 0.0000010$

Melalui perbandingan hasil kuantitatif probabilitas posterior di atas, nilai peluang pada kelas *Bullying* terbukti jauh lebih dominan dibandingkan kelas *Non-bullying* (0.0007875 > 0.0000010). Dengan demikian, sistem secara logis menetapkan kalimat tersebut ke dalam Kelas Bullying (Label 1).

3.5 Hasil Evaluasi Performa Model Klasifikasi

Setelah seluruh 130 dokumen data uji diprediksi secara otomatis oleh model Multinomial Naive Bayes, tingkat kebenaran prediksi diukur secara ketat dengan mencocokkan hasil tebakan model terhadap label asli dokumen (*ground truth*). Hasil pemetaan visual tersebut disajikan ke dalam tabel matriks kebingungan (*Confusion Matrix*) berikut.



Gambar 4. Confusion Matrix Hasil Pengujian Sistem

Berdasarkan persebaran kuantitas numerik pada Tabel 3, metrik evaluasi keandalan performa klasifikasi sistem dihitung secara presisi menggunakan persamaan matematis standar sebagai berikut:

a) **Akurasi (*Accuracy*)**
Rasio ketepatan prediksi total model terhadap keseluruhan data uji.

Akurasi $= \frac{TP+TN}{Total} = \frac{55+58}{130} = \frac{113}{130} = 86.92\% \rightarrow \mathbf{87}\%$

b) **Presisi (*Precision*)**
Tingkat keandalan model dalam memprediksi kebenaran kelas *cyberbullying*.

Presisi Kelas Bullying $= \frac{TP}{TP+FP} = \frac{55}{55+7} = \frac{55}{62} = 88.70\% \rightarrow \mathbf{89}\%$

c) ***Recall***

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

Kemampuan sistem dalam menjaring kembali objek kasus *cyberbullying* yang sebenarnya terjadi.

$$\text{Recall Kelas Bullying} = \frac{TP}{TP+FN} = \frac{55}{55+10} = \frac{55}{65} = 84.61\% \rightarrow \mathbf{85\%}$$

d) **F1-*Score***
Nilai penyeimbang rata-rata harmonik antara metrik presisi dan recall.

$$\text{F1-Score Kelas Bullying} = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} = 2 \times \frac{0.89 \times 0.85}{0.89+0.85} = \mathbf{87\%}$$

```
Accuracy: 0.8307692307692308

Classification Report:

              precision    recall  f1-score   support

Non-bullying       0.80      0.88      0.84        65
    Bullying       0.86      0.78      0.82        65

    accuracy                           0.83       130
   macro avg       0.83      0.83      0.83       130
weighted avg       0.83      0.83      0.83       130
```

Gambar 5. Tampilan Classification Report Output Sistem

Meskipun model menghasilkan performa yang tinggi dengan akurasi 87%, dilakukan analisis lebih mendalam terhadap adanya galat prediksi berupa 7 kasus *False Positive* (FP) dan 10 kasus *False Negative* (FN).

Kasus *False Positive* terjadi ketika komentar *non-bullying* salah diprediksi sebagai *bullying*. Hal ini disebabkan oleh kelemahan mendasar algoritma Naive Bayes yang mengasumsikan setiap kata berdiri sendiri (*conditional independence*). Ketika seorang pengguna menuliskan komentar pujian informal yang menggunakan kata penegas berupa kata umpatan akrab (misalnya: *"anjing keren banget lu bro"*), model secara bias langsung mengklasifikasikannya sebagai *bullying*. Hal ini terjadi karena fitur kata "anjing" memiliki bobot probabilitas *likelihood* yang terlampau tinggi pada kelas *bullying* berdasarkan data latih, sehingga perkalian probabilitas independennya mengabaikan konteks positif di sekitarnya.

Sementara itu, kasus *False Negative* terjadi pada komentar perundungan yang gagal diidentifikasi oleh sistem karena penggunaan gaya bahasa sarkasme atau sindiran halus yang tidak memuat kata-kata kasar vulgar secara langsung. Karena korpus TF-IDF memecah teks berbasis frekuensi kemunculan kata, struktur kalimat sindiran yang menggunakan kata-kata netral atau tampak sopan secara algoritmik akan memiliki nilai kedekatan peluang yang lebih tinggi pada kelas *non-bullying*, sehingga sistem meloloskan komentar perundungan tersebut.

Secara keseluruhan, pencapaian nilai akurasi hingga mencapai 87% ini menegaskan bahwa pemodelan klasifikasi menggunakan algoritma Multinomial Naive Bayes yang dikombinasikan dengan ekstraksi pembobotan fitur gabungan Unigram dan Bigram terbukti sangat sukses dan andal untuk diimplementasikan pada pemrosesan teks ulasan media sosial.

4. KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan dan mengevaluasi sistem deteksi otomatis untuk ujaran perundungan siber (*cyberbullying*) pada kolom komentar media sosial Instagram menggunakan algoritma Multinomial Naive Bayes. Melalui arsitektur pembersihan data (*text preprocessing*) yang memadukan *case folding*, normalisasi kata tidak baku (*slang word*), *cleansing*, *tokenizing*, dan *stopword removal*, gangguan teks (*noise*) yang menjadi karakteristik utama bahasa media sosial berhasil direduksi secara signifikan. Penerapan ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF) yang dikombinasikan dengan pendekatan *N-Gram* (gabungan *Unigram* dan *Bigram*) terbukti sangat efektif dalam menangkap konteks frasa utuh dari 650 dokumen data yang dievaluasi.

Berdasarkan pengujian klasifikasi dengan skenario pembagian data latih dan data uji sebesar 80:20, model komputasi yang dibangun mampu menghasilkan performa yang tangguh dan stabil. Hasil pemetaan *Confusion Matrix* membuktikan bahwa sistem ini mencapai tingkat Akurasi sebesar

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 4 Tahun 2026
ISSN 3025-0919 (media online)
Hal 965-974

87%, Presisi 89%, Recall 85%, dan F1-Score 87%. Analisis frekuensi kata juga mengonfirmasi bahwa kata-kata makian kasar secara konsisten menjadi indikator terkuat pada kelas perundungan. Secara keseluruhan, pemodelan ini menghasilkan draf sistem deteksi yang optimal, akurat, dan dapat diandalkan. Penelitian ini diharapkan dapat menjadi landasan ilmiah dalam pengembangan fitur moderasi otomatis pada platform digital guna mencegah meluasnya dampak negatif perundungan siber.

Meskipun sistem klasifikasi yang dibangun telah berjalan dengan efisien dan presisi, terdapat beberapa ruang pengembangan untuk penelitian selanjutnya. Direkomendasikan untuk memperbesar skala pengumpulan data komentar lintas platform (seperti TikTok atau X/Twitter) guna menguji stabilitas model pada volume korpus teks yang lebih masif dan beragam. Selain itu, eksplorasi terhadap algoritma *Deep Learning* atau penambahan fitur analisis kontekstual tingkat lanjut sangat disarankan untuk mengatasi batasan sistem saat ini, terutama dalam mengenali sentimen perundungan yang dibungkus dengan gaya bahasa sarkasme atau sindiran halus yang tidak menggunakan kata kasar secara langsung

REFERENCES

Baehaqi, F., & Cahyono, N. (2024). Analisis sentimen terhadap cyberbullying pada komentar di Instagram menggunakan algoritma Naïve Bayes. *Indonesian Journal of Computer Science, 13*(1), 1051–1063.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). *Detection of cyberbullying incidents on the Instagram social network*. Retrieved from http://arxiv.org/abs/1503.03909

Rachmat, A., & Lukito, Y. (2017). Deteksi komentar spam bahasa Indonesia pada Instagram menggunakan Naive Bayes. *ULTIMATICS, 9*(1).

Ruziqiana, M. F., Hidayah, L., & Rasyidi, M. A. (2024). Detection of cyberbullying using SVM, Naive Bayes, and Random Forest algorithm. *Jurnal Informatika dan Teknik Elektro Terapan, 12*(3). https://doi.org/10.23960/jitet.v12i3.5283