Penerapan Term Frequency-Inverse Document Frequency Pada Klasifikasi Komentar *Cyberbullying* pada Platform Instagram

Dede Brahma Arianto[1*], Iqbal Fernando[2], Noviyanti[3]

Program Studi Informatika, Fakultas Sains dan Teknik, Universitas Faletehan.
Jl. Raya Cilegon KM. 06, Kramatwatu, Serang, Banten 42182, Indonesia.
Email : dedebrahma@uf.ac.id[1], iqbal.28nando@gmail.com[2],
noviantinovi055@gmail.com[3]

Artikel Histori:
Disubmit: Mei 2026
Diterima: Juni 2026
Diterbitkan: Juni 2026

DOI
10.33005/jifti.v8i1.221

ABSTRAK

Interaksi pada kolom komentar Instagram dapat memuat pesan yang mengarah pada cyberbullying sehingga diperlukan mekanisme otomatis untuk membantu proses identifikasinya. Penelitian ini membandingkan Naïve Bayes, Decision Tree, dan K-Nearest Neighbor dalam mengklasifikasikan 650 komentar Instagram berbahasa Indonesia. Teks dinormalisasi dan direpresentasikan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF), kemudian ketiga model dievaluasi menggunakan 5-fold cross-validation. Hasil eksperimen menunjukkan bahwa Naïve Bayes memperoleh kinerja tertinggi dengan accuracy 84%, precision 84%, recall 85%, dan F1-score 84%. K-NN menghasilkan accuracy 79%, sedangkan Decision Tree mencapai 73%. Confusion matrix memperlihatkan bahwa Naïve Bayes menghasilkan distribusi kesalahan yang relatif seimbang antara kelas bullying dan non-bullying. Pada konfigurasi data dan representasi fitur yang digunakan, pendekatan probabilistik Naïve Bayes memberikan hasil yang lebih baik dibandingkan algoritma berbasis pohon maupun kedekatan jarak.

Kata Kunci: cyberbullying, klasifikasi, instagram, machine learning, TF-IDF.

PENDAHULUAN

Intensitas penggunaan media sosial membuat kolom komentar menjadi salah satu ruang interaksi digital yang sangat aktif. Instagram menyediakan fasilitas tersebut untuk mendukung komunikasi antarpengguna, tetapi pada saat yang sama dapat menjadi media munculnya komentar yang mengarah pada cyberbullying (Rahmanda & Setiawan, 2024). Bentuknya dapat berupa penghinaan, serangan personal, pelecehan verbal, atau ungkapan kebencian yang berpotensi memberikan dampak psikologis kepada korban (Sri Dianti, 2024). Volume komentar yang besar menyebabkan proses pemeriksaan secara manual menjadi sulit dilakukan secara konsisten.

Pendekatan komputasional dapat digunakan untuk membantu memilah komentar berdasarkan pola yang terdapat pada data berlabel. Dalam penelitian ini, komentar direpresentasikan menggunakan TF-IDF sehingga setiap teks dapat diproses dalam bentuk vektor numerik. Representasi yang sama kemudian diberikan kepada Naïve Bayes, Decision Tree, dan K-Nearest Neighbor (K-NN). Penggunaan fitur yang seragam memungkinkan perbedaan hasil lebih merefleksikan karakteristik algoritma yang dibandingkan, bukan perbedaan perlakuan terhadap data.

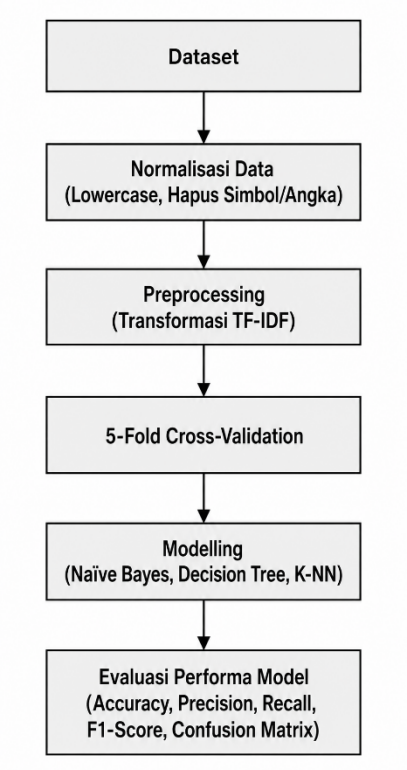
Hutagalung et al. (2021) telah menerapkan Naïve Bayes untuk mendeteksi cyberbullying pada komentar Instagram. Penelitian tersebut menunjukkan potensi metode probabilistik dalam menangani klasifikasi komentar, tetapi belum membandingkannya dengan beberapa model pada konfigurasi data yang sama. Manoppo dan Fudholi (2021) membandingkan Naïve Bayes dan Support Vector Machine dalam identifikasi cyberbullying berdasarkan unsur tindak pidana. Studi tersebut memiliki fokus berbeda dan belum mengevaluasi Naïve Bayes, Decision Tree, dan K-NN secara bersamaan pada komentar Instagram berbahasa Indonesia.

Penelitian ini mengisi celah tersebut melalui perbandingan tiga algoritma menggunakan dataset, tahapan praproses, dan representasi TF-IDF yang sama. TF-IDF dikonfigurasi dengan max_features=3000, min_df=3, dan max_df=0.8 untuk mengendalikan ruang fitur sebelum proses klasifikasi. Evaluasi dilakukan dengan accuracy, precision, recall, F1-score, dan confusion matrix.

Tujuan penelitian adalah mengidentifikasi algoritma yang memberikan kinerja terbaik untuk klasifikasi komentar bullying dan non-bullying pada dataset yang digunakan. Kontribusi penelitian terletak pada evaluasi langsung ketiga pendekatan dalam kondisi eksperimen yang seragam sehingga hasilnya dapat digunakan sebagai referensi awal dalam pengembangan sistem penyaringan komentar otomatis.

METODE PENELITIAN

Eksperimen dirancang sebagai perbandingan tiga model klasifikasi pada satu skema pemrosesan teks yang identik. Alur analisis dimulai dari penyiapan dataset, normalisasi komentar, pembentukan fitur TF-IDF, validasi lima fold, pemodelan, hingga pengukuran performa. Rangkaian proses tersebut dirangkum pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

a. Pengumpulan Data

Dataset terdiri atas 650 komentar Instagram berbahasa Indonesia yang diperoleh dari sumber open data. Setiap komentar telah memiliki kelas *bullying* atau *non-bullying* sebagai target klasifikasi. Sumber asli dataset perlu dicantumkan secara eksplisit agar asal data dapat ditelusuri.

b. Normalisasi Data

Eksperimen menggunakan atribut komentar sebagai masukan dan kategori sebagai target. Label bullying dikodekan sebagai 1, sedangkan non-bullying sebagai 0. Teks kemudian diseragamkan ke huruf kecil dan dibersihkan dari angka, simbol, serta karakter yang tidak digunakan dalam pembentukan fitur.

c. Preprocessing

Representasi numerik dibentuk menggunakan TF-IDF dengan max_features=3000, min_df=3, max_df=0.8, dan stopwords Bahasa Indonesia. Konfigurasi tersebut menghasilkan fitur yang selanjutnya diberikan secara identik kepada ketiga algoritma. Dengan cara ini, perbandingan model dilakukan menggunakan ruang fitur yang sama.

d. Skema Validasi

Validasi dilakukan dalam lima putaran menggunakan skema 5-fold CV. Pada setiap putaran, satu bagian data menjadi set evaluasi dan empat bagian lainnya membentuk model. Pergantian dilakukan sampai seluruh bagian memperoleh

giliran sebagai data evaluasi, sehingga setiap komentar berkontribusi pada pengukuran performa.

e. Modelling

Tiga algoritma digunakan sebagai model pembanding. Naïve Bayes menghasilkan prediksi berdasarkan kecenderungan probabilitas fitur TF-IDF pada masing-masing kelas. Decision Tree membangun aturan pemisahan kelas menggunakan struktur pohon dengan kriteria entropy. Sementara itu, K-NN menentukan kelas berdasarkan lima tetangga terdekat (k=5) menggunakan jarak Euclidean pada ruang fitur TF-IDF.

f. Evaluasi Performa Model

Keluaran seluruh fold digabungkan untuk memperoleh empat indikator kinerja, yaitu accuracy, precision, recall, dan F1-score. Pola kesalahan klasifikasi dianalisis melalui nilai TP, TN, FP, dan FN pada matriks konfusi.

1) Akurasi (Accuracy):

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Presisi (Precision):

$$Presisi = \frac{TP}{TP + FP}$$

3) Sensitivitas (Recall):

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

4) F1-Score:

$$F1 - Score = 2x \frac{Presisi\ x\ Recall}{Presisi\ +\ Recall}$$

HASIL DAN PEMBAHASAN

Analisis difokuskan pada perubahan data setelah praproses dan perbedaan keluaran ketiga model. Hasil disajikan mulai dari karakteristik fitur TF-IDF hingga evaluasi prediksi Naïve Bayes, Decision Tree, dan K-NN.

1. Dataset

Dataset memuat 650 komentar Instagram berbahasa Indonesia yang telah dikelompokkan ke dalam kelas bullying dan non-bullying. Atribut utama yang digunakan dalam eksperimen adalah teks komentar sebagai masukan dan kategori sebagai target klasifikasi. Distribusi data tersebut selanjutnya diproses melalui tahap normalisasi dan pembentukan fitur.

2. Normalisasi Data

Normalisasi mengubah bentuk teks tanpa mengubah label kelas. Sebagai contoh, simbol, mention, angka, dan karakter tertentu dihilangkan sebelum komentar diteruskan ke tahap pembentukan fitur seperti pada tabel 1. Label kategorikal juga dikonversi menjadi nilai numerik agar dapat diproses oleh model.

Tabel 1. Data Sebelum Dinormalisasikan

Komentar	Label	Label
"Kaka tidur yaa, udah pagi, gaboleh capek2"	Non-bullying	0
"Manusia apa bidadari sih herann deh cantik terus 😊❤️"	Non-bullying	0
"@ayu.kinantii isyan skrg berubah ya:(baju nya nakal"	Bullying	1

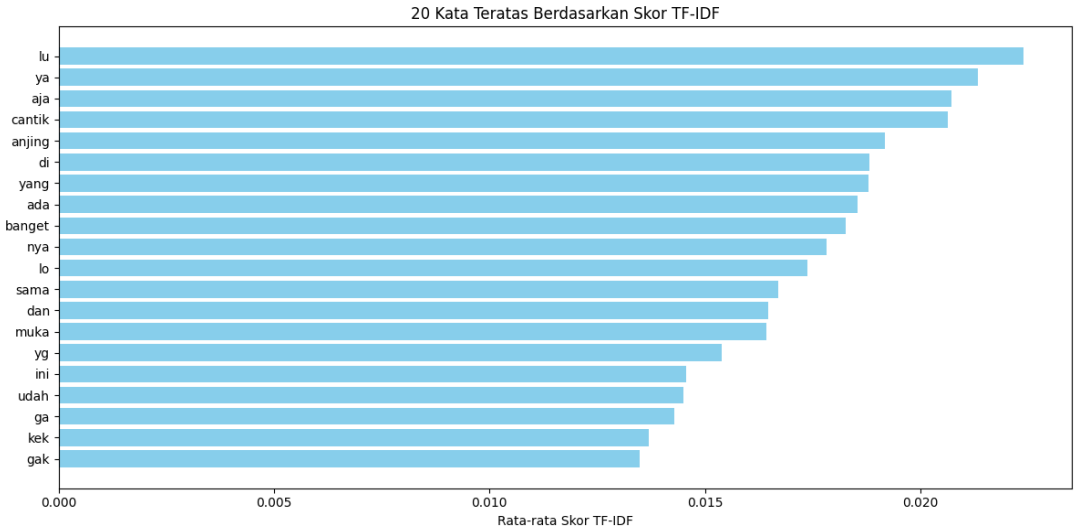
3. Preprocessing

Setelah normalisasi, korpus dipetakan ke ruang fitur TF-IDF. Konfigurasi yang digunakan menghasilkan 2.162 fitur dari 650 komentar, dan vektor inilah yang menjadi masukan bagi seluruh model.

Tabel 2. Top 10 Kata TF-IDF

No.	Kata
1	aa
2	aaa
3	aamiin
4	abaikan
5	abalabal
6	abatartila
7	abis
8	abu
9	activity
10	ad

Penerapan konfigurasi TF-IDF menghasilkan matriks berukuran 650 × 2.162. Artinya, 2.162 fitur memenuhi kriteria penyaringan dan digunakan untuk merepresentasikan seluruh komentar. Analisis bobot rata-rata memperlihatkan sejumlah kata yang memiliki kontribusi relatif tinggi pada korpus. Distribusi 20 fitur dengan skor tertinggi disajikan pada Gambar 2.



Gambar 2. Kata Teratas Berdasarkan Skor TF - IDF

Kata seperti “lu”, “ya”, “aja”, “cantik”, dan “anjing” memperoleh bobot relatif tinggi pada korpus. Kemunculannya yang tidak merata antar komentar membuat kata-kata tersebut memperoleh nilai TF-IDF lebih besar dibandingkan sejumlah fitur lainnya.

4. Evaluasi Model

Berikut adalah penjelasan evaluasi hasil model klasifikasi komentar *cyberbullying* berdasarkan output.

a) *Naïve Bayes*

Naïve Bayes memperoleh rata-rata accuracy sekitar 84%. Precision dan recall kedua kelas berada pada rentang yang berdekatan sehingga kesalahan prediksi tidak terkonsentrasi secara kuat pada salah satu kelas.

```
=== Naïve Bayes ===
Akurasi tiap fold : [0.8615 0.8462 0.8615 0.8    0.8538]
Rata-rata Akurasi (5-Fold CV): 0.8446 (+/- 0.0230)
              precision    recall  f1-score   support

           0       0.85      0.84      0.84       325
           1       0.84      0.85      0.84       325

    accuracy                           0.84       650
   macro avg       0.84      0.84      0.84       650
weighted avg       0.84      0.84      0.84       650
```

Gambar 3. Hasil Evaluasi Model *Naïve Bayes*

b) *Decision Tree*

Decision Tree menghasilkan rata-rata accuracy sekitar 73%, terendah di antara ketiga algoritma. Recall kelas bullying sebesar 0,70 menunjukkan bahwa sebagian data positif masih gagal dikenali oleh model.

```
=== Decision Tree ===
Akurasi tiap fold : [0.7615 0.7077 0.7308 0.6923 0.7615]
Rata-rata Akurasi (5-Fold CV): 0.7308 (+/- 0.0279)
              precision    recall  f1-score   support

           0       0.72      0.76      0.74       325
           1       0.74      0.70      0.72       325

    accuracy                           0.73       650
   macro avg       0.73      0.73      0.73       650
weighted avg       0.73      0.73      0.73       650
```

Gambar 4. Hasil Evaluasi Model *Decision Tree*

c) *K-Nearest Neighbor* (K-NN)

K-NN mencapai rata-rata accuracy sekitar 79%. Pada kelas bullying, model memperoleh recall 0,81 dan precision 0,78 sehingga hasilnya berada di antara Naïve Bayes dan Decision Tree.

```
=== K-NN ===
Akurasi tiap fold : [0.8    0.7462 0.7923 0.7769 0.8154]
Rata-rata Akurasi (5-Fold CV): 0.7862 (+/- 0.0235)
              precision    recall  f1-score   support

           0       0.80      0.77      0.78       325
           1       0.78      0.81      0.79       325

    accuracy                           0.79       650
   macro avg       0.79      0.79      0.79       650
weighted avg       0.79      0.79      0.79       650
```

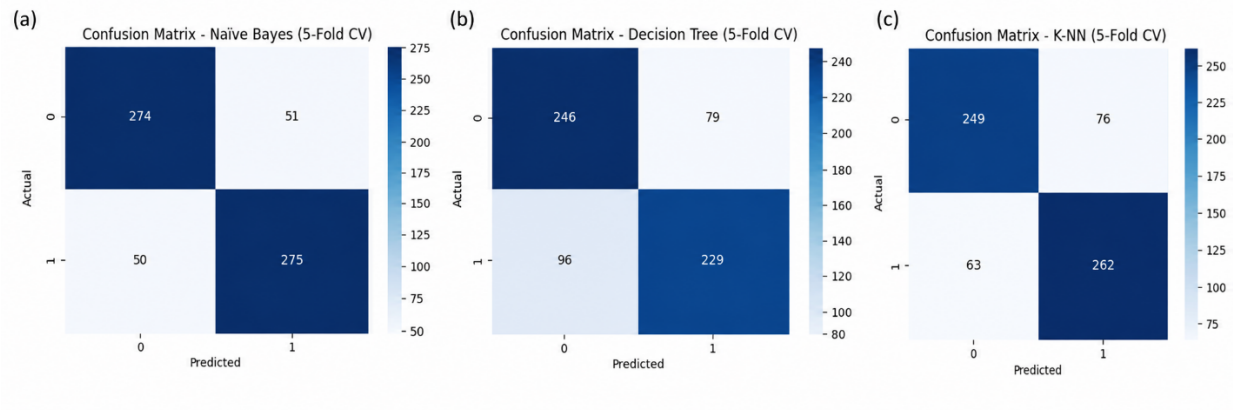
Gambar 5. Hasil Evaluasi Model K-NN

5. Hasil Perbandingan

Tabel 3 memperlihatkan urutan performa yang konsisten pada empat metrik evaluasi. Naïve Bayes berada pada posisi pertama dengan accuracy dan F1-score masing-masing 0,84, diikuti K-NN sebesar 0,79, sedangkan Decision Tree memperoleh accuracy 0,73 dan F1-score 0,72. Perbedaan tersebut menunjukkan bahwa, pada representasi TF-IDF yang digunakan, Naïve Bayes menghasilkan prediksi paling stabil di antara tiga model yang diuji. Distribusi kesalahan masing-masing model dianalisis lebih lanjut melalui matriks konfusi pada Gambar 6.

Tabel 3. Perbandingan Performa Model

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naïve Bayes	**0.84**	**0.84**	**0.85**	**0.84**
Decision Tree	0.73	0.74	0.70	0.72
K-NN	0.79	0.78	0.81	0.79



Gambar 6. Perbandingan Confusion Matrix: (a) Naïve Bayes, (b) Decision Tree, dan (c) K-NN

Confusion matrix memperlihatkan perbedaan jumlah prediksi benar antar model. Naïve Bayes menghasilkan 549 prediksi benar, terdiri atas 274 true negative dan 275 true positive, dengan 51 false positive dan 50 false negative. K-NN menghasilkan 511 prediksi benar, sedangkan Decision Tree menghasilkan 475 prediksi benar. Temuan tersebut konsisten dengan hasil metrik agregat yang menempatkan Naïve Bayes sebagai model dengan performa tertinggi.

SIMPULAN

Pengujian lima fold menunjukkan bahwa pendekatan probabilistik memberikan hasil paling baik pada fitur TF-IDF yang dibentuk dari dataset penelitian. Naïve Bayes menempati peringkat pertama dengan accuracy 0,84, diikuti K-NN 0,79 dan Decision Tree 0,73. Pola pada matriks konfusi juga menunjukkan bahwa Naïve Bayes memiliki distribusi kesalahan yang paling seimbang dibandingkan dua model lainnya.

Cakupan data masih terbatas pada 650 komentar dan belum merepresentasikan variasi bahasa media sosial secara luas. Eksperimen berikutnya dapat memperluas korpus serta menggunakan representasi kontekstual seperti IndoBERT untuk menguji apakah informasi semantik dapat meningkatkan kemampuan identifikasi cyberbullying.

DAFTAR PUSTAKA

Saputri, F., et al. (2020). Klasifikasi *Hate Speech* pada *Twitter* Menggunakan Random Forest. *Jurnal Teknologi Informasi*, 12(1).

Dimas Bayu (2022) *APJII: Pengguna Internet Indonesia Tembus 210 Juta pada 2022*, *https://dataindonesia.id/digital/detail/apj ii-pengguna-internet-indonesia-tembus- 210-juta-pada-2022*.

Ramadhani, M. & Sasmi, D. (2021). Deteksi Ujaran Kebencian Berbahasa Indonesia Menggunakan TF-IDF dan SVM. *Jurnal Ilmiah Komputer dan Informatika (JIKI)*, 9(2).

Hadiyat, A. & Siregar, F. (2022). Perbandingan Kinerja Naïve Bayes dan K-NN dalam Klasifikasi Komentar YouTube. *Jurnal RESTI*, 6(3).

Hutagalung, A. S., Putra Negara, A. B., & Pratama, E. E. (2021). *Aplikasi Pendeteksi Cyberbullying Terhadap Komentar Postingan Media Sosial Instagram dengan Metode Naïve Bayes Classifier Berbasis Website*. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 9(3).

Kamal, M. M., & Rahayu, Y. (2021). Klasifikasi Komentar Cyberbullying pada Platform Instagram Menggunakan Algoritma Naïve Bayes, Decision Tree, dan K-NN. *Jurnal Ilmiah Informatika KOMPUTA, 10*(2), 115-123.

Manoppo, T.N. and Fudholi, D.H. (2021) 'Deteksi Cyberbullying berdasarkan Unsur Perbuatan Pidana yang Dilanggar dengan Naive Bayes dan Support Vector Machine', *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 5(1), pp. 10 - 19.

Qonita, N. S., & Ramadhani, R. (2021). Text Classification untuk Deteksi Cyberbullying pada Media Sosial Menggunakan Metode Machine Learning. *Jurnal Teknologi dan Informatika*, 13(2), 109–117.

Iqbal, M., Davy Wiranata, A., Suwito, R., & Faiz Ananda, R. (2023a). KLIK: Kajian Ilmiah Informatika dan Komputer Perbandingan Algoritma Naïve Bayes, KNN, dan Decision Tree terhadap Ulasan Aplikasi Threads dan Twitter. *Media Online*, *4*(3), 1799–1807.

Sari, N. L., & Muslim, R. (2023). Implementasi Algoritma Machine Learning dalam Klasifikasi Ujaran Kebencian di Media Sosial. *Jurnal Teknologi dan Sistem Komputer*, 11(1), 23–30.

Normalita, A., & Arianto, D. B. (2025). Analisis Sentimen Dengan Metode Naïve Bayes Terhadap UlasanHotel New Saphir Yogyakarta Pada Platform Tripadvisor. Jurnal Teknologi Informasi, 1(1), 26-33.