Jurnal Informatika: Jurnal pengembangan IT

Vol. 9, No. 3, 2024
ISSN : 2477-5126, DOI:10.30591/jpit.v9i3.6611

Analisis Sentimen Inses di Social Media menggunakan Algoritma Naïve Bayes

Tasya Salsabilla [1], Debby Alita [2]
[1,2]Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia,
[1,2]Jl.ZA. Pagar Alam No.9-11, Kota Bandar Lampung, Lampung 35132, Indonesia

Info Artikel

Riwayat Artikel:

Received 2024-03-06
Revised 2024-12-13
Accepted 2024-12-14

Corresponding Author:

Suttichai Premrudeeprechacharn
Email: suttichai@mail.com

Abstract *– Sexual violence, especially against women and children, is a serious problem in Indonesia. Cases are increasing every year, including incest, which involves sexual relations between close family members. Girls, who are often considered weak and vulnerable, are the main victims. The latest data from the National Commission on Violence Against Women records a decrease in incest cases from 1,210 in 2017 to 215 in 2020. However, attention is still needed, especially because biological fathers are the largest perpetrators. This research uses the Naïve Bayes algorithm for sentiment analysis. This algorithm is an effective classification method based on Bayes' theorem with simple assumptions but is quite effective. Assuming that each feature in the data is independent, Naïve Bayes can work well in text analysis. The research results showed an accuracy rate of 94%. Continued attention to sexual violence, especially incest, is needed to protect vulnerable girls. Protection efforts must continue to be improved, including the application of sentiment analysis methods such as Naïve Bayes for monitoring and early detection. Public awareness and cross-sector cooperation are also key in overcoming this phenomenon.*

Abstrak *– Kekerasan seksual, terutama terhadap perempuan dan anak-anak, menjadi permasalahan serius di Indonesia. Kasus-kasus semakin meningkat setiap tahun, termasuk inses, yang melibatkan hubungan seksual di antara anggota keluarga dekat. Anak perempuan, yang seringkali dianggap lemah dan rentan, menjadi korban utama. Data terbaru dari Komnas Perempuan mencatat penurunan kasus inses dari 1.210 pada 2017 menjadi 215 pada 2020. Namun, perhatian tetap diperlukan, terutama karena ayah kandung merupakan pelaku terbanyak. Penelitian ini menggunakan algoritma Naïve Bayes untuk analisis sentimen. Algoritma ini adalah metode klasifikasi efektif berdasarkan teorema Bayes dengan asumsi sederhana namun cukup efektif. Dengan asumsi bahwa setiap fitur dalam data adalah independen, Naïve Bayes dapat bekerja dengan baik dalam analisis teks. Hasil penelitian menunjukkan tingkat akurasi sebesar 94%. Pentingnya perhatian terus-menerus terhadap kekerasan seksual, khususnya inses, dibutuhkan untuk melindungi anak perempuan yang rentan. Upaya perlindungan harus terus ditingkatkan, termasuk penerapan metode analisis sentimen seperti Naïve Bayes untuk pemantauan dan deteksi dini. Kesadaran masyarakat dan kerja sama lintas sektor juga menjadi kunci dalam mengatasi fenomena ini.*

Kata Kunci*: Algoritma Naïve Bayes, Analisis Sentimen, Inses, Kekerasan Seksual.*

I. PENDAHULUAN

Kekerasan seksual terutama pada wanita dan anak-anak, jadi salah satu fenomena yang sangat mengkhawatirkan di Indonesia. Setiap tahunnya, kasus-kasus kekerasan kian sering terjadi, menyebar di hampir seluruh provinsi. Salah satu bentuk kekerasan seksual yang sangat mengkhawatirkan adalah inses, yaitu hubungan seksual antara anggota keluarga dekat, seperti antara orangtua dan anak perempuan, ibu dengan anak laki-laki, atau antara saudara-saudara. Anak perempuan adalah kelompok yang sangat rentan terhadap kekerasan seksual karena sering dianggap lemah dan bergantung pada orang dewasa di sekitar mereka [1]. Berdasarkan data Komnas Perempuan selama 4 tahun terakhir hingga tahun 2021, kasus inses mengalami fluktuasi. Pada tahun 2017 tercatat 1.210 kasus, turun menjadi 1.017 kasus pada tahun 2018, kemudian menurun lagi menjadi 822 kasus pada tahun 2019. Pada tahun 2020, jumlahnya menurun drastis menjadi 215 kasus, namun perlu tetap diperhatikan. Fokus perhatian sebaiknya diberikan terutama pada kasus yang melibatkan ayah kandung, yang merupakan pelaku terbanyak sebanyak 425 orang [2].

Inses pada umumnya merujuk pada hubungan seksual di antara individu yang memiliki hubungan keluarga dekat atau ikatan darah, yang dianggap melanggar norma adat, hukum, dan agama. Ini melibatkan tiga situasi utama, yaitu inses orangtua-anak (hubungan antara orang tua dan anak), inses saudara (hubungan antara saudara kandung), dan inses keluarga (hubungan seksual di antara anggota keluarga dekat). Sosial media termasuk platform *twitter*, telah menjadi sarana bagi masyarakat untuk berbagi pemikiran, pendapat, dan pengalaman mereka secara luas [3]. Melalui postingan dan percakapan di *twitter*, masyarakat dapat memperoleh dan menyebarkan informasi tentang berbagai fenomena, termasuk masalah sensitif seperti inses [4]. Dengan memanfaatkan data yang tersedia di media sosial

khususnya *twitter*, peneliti dapat melakukan analisis terhadap pendapat dan opini masyarakat terkait fenomena inses [5].

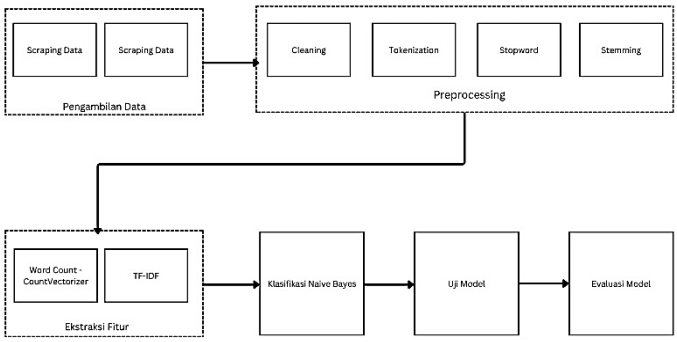
Salah satu cara untuk menganalisis perasaan atau pendapat seseorang adalah menggunakan algoritma *Naïve Bayes*. Algoritma ini merupakan metode pengelompokan yang berdasarkan teorema *Bayes* dengan asumsi sederhana namun efektif. Asumsi pokok dari algoritma ini adalah bahwa setiap bagian informasi dalam data tidak saling bergantung satu sama lain, meski sebenarnya dalam keadaan nyata hal tersebut mungkin tidak sepenuhnya berlaku [6]. Meskipun demikian, asumsi ini memungkinkan algoritma *Naïve Bayes* untuk bekerja dengan baik dalam berbagai konteks, termasuk dalam analisis teks dan klasifikasi dokumen. Proses kerja algoritma *Naïve Bayes* dimulai dengan menghitung probabilitas dari masing-masing kelas yang ada dalam data, berdasarkan frekuensi kemunculan fitur-fitur tertentu [7]. Algoritma ini dikenal karena kemampuannya dalam melakukan klasifikasi dengan baik berdasarkan probabilitas, sehingga cocok digunakan untuk menganalisis data opini dan pendapat yang terdapat dalam postingan di *twitter* mengenai inses [8].

Penelitian ini bertujuan untuk menganalisis opini masyarakat terkait dengan fenomena inses. Penelitian ini juga memberikan kontribusi dalam upaya pencegahan kekerasan seksual, khususnya inses, dengan menyajikan hasil analisis yang dapat digunakan sebagai dasar untuk para pemangku kebijakan.

Berdasarkan serangkaian penelitian sebelumnya terdapat beberapa temuan penting terkait penerapan algoritma *Naïve Bayes* dalam analisis sentimen menggunakan data dari media sosial, khususnya *twitter* [9]. Menemukan bahwa tingkat akurasi sistem dalam klasifikasi data *twitter* menggunakan algoritma *Naïve Bayes* mencapai 78% dengan metode *confusion matrix*, dan meningkat menjadi 80% dengan menggunakan *k-fold cross validation*. Penelitian lain oleh [10] menunjukkan bahwa pengguna twitter di Indonesia cenderung memberikan komentar yang netral terhadap kasus anti LGBT, dengan akurasi analisis sentimen menggunakan *Naïve Bayes Classifier* sebesar 76.84%. Selanjutnya penelitian oleh [11] menyoroti pendapat masyarakat di *twitter* mengenai pemindahan Ibu Kota Negara, menemukan bahwa metode *Naïve Bayes* mampu mencapai akurasi sebesar 100% dalam *split* data dan 90.84% dalam *cross-10k folds validation*. [12] dalam penelitiannya, mengeksplorasi sentimen terhadap isu penundaan pemilu sebelum dan sesudah diutarakan oleh Muhaimin Iskandar, menemukan tingkat akurasi yang tinggi untuk analisis sentimen menggunakan *Naïve Bayes Classifier*. Terakhir, [13] menemukan bahwa algoritma *Naïve Bayes Classifier* dengan seleksi fitur *Chi Squared Statistic* berbasis *forward selection* memiliki akurasi yang lebih tinggi dibandingkan dengan metode seleksi fitur *Particle Swarm Optimization* (PSO), dengan selisih akurasi sebesar 3,11%. Berdasarkan temuan-temuan tersebut, penelitian ini akan mengeksplorasi Analisis Sentimen Inses di media sosial menggunakan algoritma *Naïve Bayes*, memperluas pemahaman tentang pola-pola sentimen masyarakat terhadap isu sensitif ini.

II. METODE

Penelitian ini menggunakan cara mengumpulkan data dengan melihat *tweet* mengenai kasus inses di *twitter*, dan metode penelitian yang dipilih adalah metode kuantitatif. Adapun tahapan penelitian yang digunakan dapat ditinjau pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Preprocessing

Preprocessing adalah serangkaian tindakan yang dilakukan untuk membersihkan, mengintegrasikan, mengubah, dan mengurangi data. Langkah ini sangat penting karena analisis data yang belum ditangani dapat menghasilkan informasi yang salah. Kualitas data akan menurun jika ada banyak informasi yang tidak relevan atau *noise* dalam data. Untuk menghasilkan data berkualitas tinggi, sangat penting untuk memahami dan menerapkan dengan benar teknik *pre-processing* yang digunakan untuk setiap studi kasus. Langkah *pre-processing* yang dijalankan ialah *cleansing*, *case folding*, *tokenizing*, *stopword*, dan *stemming*.

1. *Cleansing*: Proses ini digunakan untuk menghilangkan hal-hal yang tidak perlu dalam data, seperti angka, simbol, tanda baca, *emoticon*, URL, *mention*, dan sejenisnya. Tujuannya adalah membuat data jadi lebih bersih.
2. *Case Folding*: Proses ini membuat semua huruf besar diubah menjadi huruf kecil, termasuk pada awal kalimat, nama orang, nama kota, dan sebagainya.
3. *Tokenizing*: Proses ini mengubah kalimat menjadi kata, memberikan bobot pada kata, yang dibutuhkan untuk proses TF-IDF.
4. *Stopword*: Langkah keempat adalah menghapus *stopword*. Dalam proses ini, kata-kata yang dienkripsi disaring. Kata-kata yang tidak memiliki makna akan dihapus atau dihilangkan. Ini akan sangat berguna untuk menganalisis perasaan atau pendapat orang.
5. *Stemming*: *Stemming* merupakan langkah untuk merubah setiap kata dengan imbuhan menjadi bentuk dasarnya. Ini termasuk imbuhan pada awal (*prefixes*), tengah (*infixes*), dan akhir kata (*suffixes*).

B. TF-IDF

Sebelum memakai pendekatan *deep learning* atau *machine learning* untuk komputasi, studi analisis sentiment ini harus melakukan pembobotan kata. Ini memastikan bahwa setiap kata dapat diukur dan diberi nilai. Algoritma seperti TF-IDF menghitung nilai TF dan IDF untuk setiap dokumen dalam korpus sesuai dengan kriteria yang telah ditentukan. Algoritma ini memiliki kemampuan untuk melakukan pembobotan kata [14] Pengertian TF-IDF sederhana untuk menghitung jumlah kata yang muncul dalam dokumen. Salah satu rumus yang sering digunakan adalah:

$$Tf(w) = \frac{frekuensi\ muncul\ kata\ w\ di\ dokumen\ d}{\text{Total kata dokumen d}} \quad (1)$$

$$IDF(w) = log_e \frac{jumlah\ total\ dokumen}{\text{total dokumen pada kata w}} \quad (2)$$

C. Naive Bayes

Setelah proses TF-IDF selesai, dilakukan klasifikasi dan pengujian model dengan menggunakan metode *Naïve Bayes*. Metode *machine learning* ini memanfaatkan model yang berfokus pada kemungkinan atau peluang, serta memiliki kemampuan klasifikasi sebanding dengan *Decission Tree* dan *Neural Network*. Kedua model tersebut terkenal karena akurasi dan kecepatannya yang tinggi [15] Persamaan yang digunakan untuk algoritma *Naïve Bayes* adalah sebagai berikut:

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

Data X merupakan kelas yang belum diketahui, sedangkan C adalah dugaan kelas untuk data tersebut. P(C|X) adalah kemungkinan dugaan kelas C setelah memperhitungkan data X. P(C) adalah kemungkinan awal untuk kelas C. P(X|C) adalah kemungkinan atau peluang munculnya data X jika kelasnya adalah C. P(X) adalah bukti atau indikasi, yaitu seberapa mungkin munculnya data tersebut.

D. Confussion Matrix

Confussion Matrix memiliki persamaan sebagai berikut karena tujuan penelitian ini adalah untuk mengukur kinerja analisis sentimen.

$$Akurasi\ \% = \frac{TP+TN}{\text{TP+TN+FP+FN}} \times 100\% \quad (4)$$

$$Presisi\ \% = \frac{TP}{\text{TP+FN}} \times 100\% \quad (5)$$

$$Recall\ \% = \frac{TP}{\text{TP+FN}} \times 100\% \quad (6)$$

$$F1 - Score\ \% = \frac{2 \times Presisi + Recall}{\text{Presisi+Recall}} \times 100\% \quad (7)$$

Jumlah data yang benar-benar positif dan berhasil diprediksi sebagai positif disebut TP. FP terjadi ketika data seharusnya negatif, tapi diprediksi sebagai positif. FN terjadi saat data yang seharusnya negatif diprediksi sebagai negatif. TN adalah ketika data negatif diprediksi dengan benar sebagai negatif.

III. HASIL DAN PEMBAHASAN

A. Crawling data

Hasil dari tahap *crawling* data, didapatkan data sebanyak 3500 data komentar terkait dengan inses dalam rentang waktu dari bulan November 2023 hingga Januari 2024 dari seluruh pengguna *twitter* yang memberikan opini terkait kasus inses. Dalam pengumpulan data ini dilakukan dengan metode *web scrapping* menggunakan link url postingan penggguna *twitter* yang dapat dilihat pada Tabel 1 dibawah ini.

TABEL 1
HASIL CRAWLING DATA

No	Tweet
1	1.66599e+18, "Ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa Inggris Udah Jelas isinya pelaku belum diketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis gais. https://t.co/hKXARMjIQ2".
2	1.6826e+18,@6undul0h Kasus perkosaan inses bisanya pulang cepat tuhan

B. Preprocessing

Setelah data selesai dilakukan proses *crawling* lalu masuk ketahap selanjutnya yaitu proses *Preprocessing* pada tahap ini dilakukan dengan beberapa teknik seperti *case folding*, *tokenize data*, *stopword removal data*, dan *stemming data* yang akan dijelaskan secara rinci pada bagian ini.

C. Case Folding

Pada tahap ini digunakan teknik *case folding* yaitu penghapusan beberapa karakter, pembersihan data rendundan, serta mengubah semua kata menjadi huruf kecil. Huruf yang tidak bermanfaat dihapus dan dianggap delimiter. Delimiter merupakan urutan satu karakter yang digunakan untuk memisahkan data. Hasil *Case Folding* dapat dilihat pada Tabel 2.

TABEL 2
HASIL CASE FOLDING

Tweet	Case Folding
1.66599e+18, "Ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa Inggris Udah Jelas isinya pelaku belum diketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis gais. https://t.co/hKXARMjIQ2".	ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa inggris udah jelas isinya pelaku belum di ketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis.
1.6826e+18,@6undul0h Kasus perkosaan inses bisanya pulang cepat tuhan	Kasus perkosaan inses bisanya pulang cepat tuhan

D. Tokenisasi

Dalam tahap tokenisasi ini berfungsi untuk memecah kata-kata pada kalimat, paragraf, atau dokumen yang menjadi token individual atau potongan kata tunggal. Tokenisasi membantu dalam analisis sentiment dengan menemukan kata-kata kunci yang menunjukkan perasaan positif atau negatif. Dengan menerapkan tokenisasi, penulis dapat mengolah kata secara terpisah, sehingga mempercepat proses analisis. Pada penelitian ini menggunakan fungsi dari *library word_tokenize* dan NLTK. Hasil dari proses tokenisasi data dapat dilihat pada Tabel 3.

TABEL 3
HASIL TOKENISASI

Tweet	Tokenizing Data
ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa inggris udah jelas isinya pelaku belum di ketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis.	'ini', 'thread', 'aku', 'anggap', 'merupakan', 'contoh', 'dari', 'kasus', 'darurat', 'membaca', 'atau', 'emang', 'sender', 'ga', 'ngerti', 'bahasa', 'inggris', 'udah', 'jelas', 'isinya','pelaku','belum', 'diketahui','siapa', 'dan', 'komplainnya', dilontarkan', 'karena', 'buku', 'nya', 'tu', 'ada', 'konten', 'inses', 'prostitusi', 'dan', 'perkosaan', 'gais'

Kasus perkosaan inses bisanya pulang cepat tuhan	‘kasus’, ‘perkosaan’, ‘inses’, ‘bisanya’, ‘pulang’, ‘cepat’, ‘tuhan’

E. Stopword Removal

Tahap *stopword removal* merupakan sebuah proses penghapusan kata yang tidak relevan atau tidak memberikan banyak informasi dalam teks. Tujuan dari proses ini untuk mengurangi munculnya kata yang tidak dianggap penting dan memudahkan proses klasifikasi. Hasil proses *Stopword Removal* dapat dilihat pada Tabel 4.

TABEL 4
HASIL STOPWORD REMOVAL

Tweet	Stopword removal
ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa inggris udah jelas isinya pelaku belum di ketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis.	'contoh', 'dari', 'kasus', 'jelas','pelaku','belum', 'diketahui','siapa', 'dan', 'komplainnya', 'ada', 'konten', 'inses', 'prostitusi', 'perkosaan'
Kasus perkosaan inses bisanya pulang cepat tuhan	‘kasus’, ‘perkosaan’, ‘inses’

F. Stemming

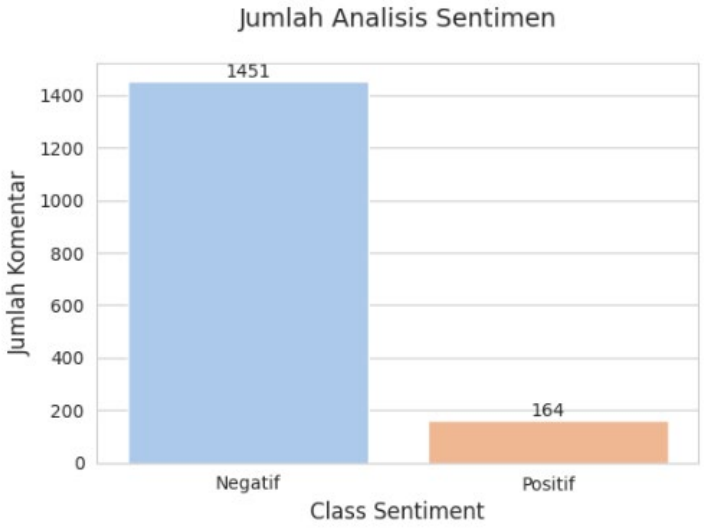
Stemming merupakan proses mengembalikan kata-kata yang berimbuhan menjadi bentuk dasarnya. Tujuannya untuk menyederhanakan kata-kata dalam teks sehingga peneliti menemukan kata-kata kunci yang terkait dengan sentimen yang diungkapkan. Pada tahap ini digunakan teknik stemming agar menyederhanakan teks ddan memperbaiki kualitas data. Hasil dari proses *stemming* dapat dilihat pada Tabel 5.

TABEL 5
HASIL STEMMING

Tweet	Stopword removal
ini thread aku anggap contoh kasus darurat membaca atau emang seder ga ngerti bahasa inggris udah jelas isinya pelaku belum di ketahui itu siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, dan perkosaan gadis.	ini thread aku anggap contoh kasus darurat membaca emang seder ngerti bahasa inggris udah jelas isinya pelaku belum di ketahui siapa dan konten dilontarkan karena bukunya isinya inses, prostitusi, perkosaan gadis.
Kasus perkosaan inses bisanya pulang cepat tuhan	Kasus perkosaan inses bisanya pulang cepat tuhan

G. Labelling Data

Setelah tahap *preprocessing* selesai, selanjutnya data yang telah dibersihkan dilakukan proses *labelling*. *Labelling* merupakan proses memberikan label pada data komentar dengan cara membagi ke dalam 2 kategori *labelling* data yaitu positif dan negatif berdasarkan sentimen yang terkandung didalamnya. Data yang didapat dari hasil *preprocessing* sebanyak 1615 data komentar. Hasil dari proses *Labelling* data dapat dilihat pada Gambar 2 dibawah ini.



Gambar 2. *wordcould* kasus inses

Setelah dilakukan proses *labelling* dan pembagian data menjadi 2 kategori yaitu label positif dan label negatif, hasil dari proses tersebut didapatkan data dengan label positif sebanyak 164 komentar, dan sebanyak 1451 komentar negative.

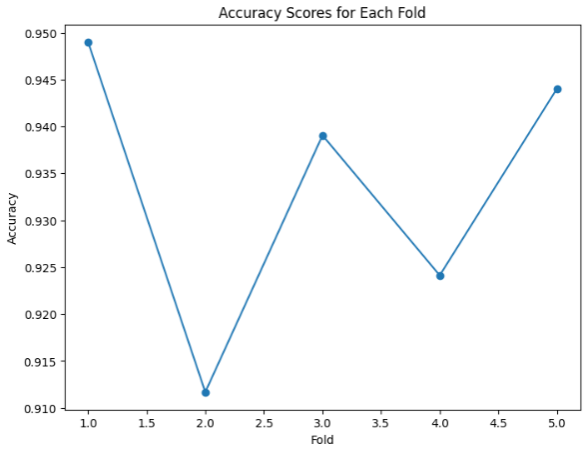
H. Pembagian data

Pada penelitian ini dilakukan 2 kali pengujian. Pertama menggunakan presentase 70% data latih dan 30% data uji, kedua menggunakan data sebanyak 80% untuk data latih dan 20% untuk data uji. Hal ini dimaksudkan untuk mencari akurasi terbaik.

Langkah selanjutnya yaitu mengimplementasikan metode *confusion matrix* dalam tahapan ini digunakan untuk memberikan informasi terkait dengan hasil prediksi klasifikasi secara aktual. Melalui *compound score* ini dapat diketahui *tweet* tersebut dikategorikan pada sentimen positif atau negatif.

I. Klasifikasi Model Naïve Bayes

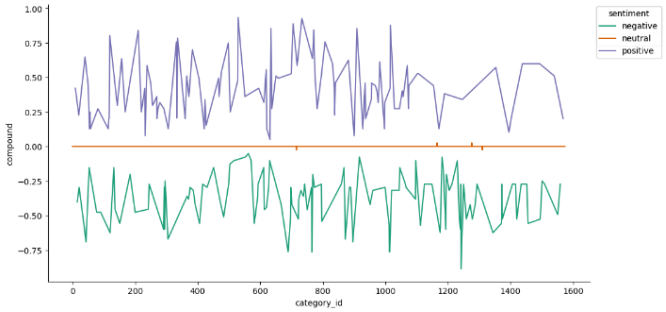
Pada tahap ini, model klasifikasi diuji dengan data yang sudah melewati tahap *preprocessing*. Pengujian ini dilakukan dengan menggunakan algoritma *naïve bayes*, pada dataset analisis sentimen opini publik terkait kasus inses agar lebih optimal lagi dalam memprediksi sentimen. Hasil dari grafik nilai akurasi dapat dilihat pada Gambar 3 di bawah ini.



Gambar 3. *Accuracy score*

Dari grafik tersebut menunjukkan skor akurasi untuk setiap *fold* dalam analisis sentimen. Dalam penelitian ini peneliti menggunakan 5 *Fold* untuk melatih performa model agar semakin tinggi skor akurasi yang dicapai. Dari grafik

tersebut, dapat dilihat bahwa skor akurasi untuk setiap *fold* berkisar antara 91% hingga 95%. Ini menunjukkan bahwa model cukup akurat dalam mengklasifikasikan data sentimen. Hasil prediksi sentimen dapat dilihat pada Gambar 4 dibawah ini.



Gambar 4. hasil prediksi sentiment

Pada Gambar 4 diatas merupakan hasil prediksi dari model *naive bayes* dimana kemampuan model dalam memprediksi pola sentimen masyarakat terhadap kasus inses dimana sangat banyak komentar negatif yang masyarakat sampaikan.

J. Evaluasi Model

Setelah selesai membuat klasifikasi model dengan algoritma *naive bayes*, langkah selanjutnya dilakukan evaluasi model dengan menggunakan proses data training dan evaluasi menggunakan teknik *imbalance oversampling*, validasi silang dan *K-Fold*. Dengan melakukan evaluasi hasil dari pelatihan model terhadap validasi data dan plot perubahan nilai akurasi pada setiap pelatihan.

1. Evaluasi akurasi model: Tabel 6 menunjukkan klasifikasi data *tweet* terkait kasus inses mengggunakan *naive bayes* menunjukkan bahwa model dengan akurasi tertinggi sebesar 94,9% adalah *classifier* yang dibuat dengan presentase pembagian data sebesar 80:20.
   Penggunaan teknik *oversampling* untuk mengatasi ketidakseimbangan kelas dan penerapan strategi *StratifiedKFold* dalam validasi silang memberikan kontribusi positif dalam meningkatkan akurasi model.

TABEL 6
PERBANDINGAN AKURASI MODEL KLASIFIKASI

Presentase pembagian	Akurasi
70:30	91.2%
80:20	94.9%

2. Teknik *Oversampling*: Teknik *oversampling* merupakan metode untuk mengatasi masalah ketidakseimbangan atau imbalance dalam dataset. Ketidakseimbangan ini terjadi karena jumlah sampel dari satu kelas jauh lebih banyak dari pada jumlah sampel kelas lainnya. Dalam penelitian ini, peneliti menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*) metode ini mensintesis sampel baru dengan menggabungkan fitur dari sampel minoritas yang ada. Keuntungan menerapkan oversampling ini adalah meningkatkan akurasi model dengan mengurangi bias terhadap kelas mayoritas, serta mengurasi resiko underfitting karena model memiliki banyak data untuk dipelajari oleh model.
3. Evaluasi *StratifiedKFold*: Metode *StrafiedKFold* merupakan pendekatan yang digunakan untuk mengukur kinerja model, metode ini membagi dataset menjadi beberapa *fold* dengan mempertahankan jumlah kelas yang sama. Tujuan utamanya untuk menghindari biar dalam evaluasi model karena ketidakseimbangan kelas. Dengan menggunakan metode *StratifiedKFold*, peneliti dapat memastikan evaluasi model dilakukan secara adil dan menghasilkan kinerja yang baik dalam proses analisis sentimen.
4. Evaluasi *recall* model: Tabel 7 menunjukkan klasifikasi data tweet terkait dengan kasus inses di media sosial menggunakan *naive bayes* menunjukkan bahwa model dengan *recall* tertinggi sebesar 100% adalah yang dibentuk dengan pembagian data sebesar 80:20.

TABEL 7
PERBANDINGAN RECALL MODEL KLASIFIKASI

Presentase pembagian	Recall
70:30	99%
80:20	100%

Teknik *oversampling* dan penggunaan *StratifiedKFold* dalam validasi menunjukkan dampak positif pada kemampuan model untuk mendeteksi kelas di kasus inses. Hasil ini secara khusus menunjukkan keefektifan

strategi untuk mengatasi ketidakseimbangan kelas dalam pengambilan sampel yang berlebihan dan validasi yang cermat dengan *StratifiedKFold* untuk meningkatkan kinerja model klasifikasi untuk teks yang terkait dengan kasus inses.

5. Evaluasi presisi model: Seperti yang ditunjukkan pada Tabel 8, peneliti menggunakan *naive bayes* untuk mengevaluasi keakuratan model dalam mengklasifikasikan data *tweet* terkait kasus inses di media sosial, dan peneliti memperoleh hasil yang signifikan. Model dengan pembagian data 80:20 mencapai presisi tertinggi sebesar 94%.

TABEL 8
PERBANDINGAN PRESISI MODEL KLASIFIKASI

Presentase pembagian	Presisi
70:30	90%
80:20	94%

Teknik *oversampling* dan penggunaan *StratifiedKFold* dalam validasi berkontribusi positif pada kemampuan model untuk memberikan hasil yang lebih akurat dan akurat ketika mengklasifikasikan data terkait kasus inses. Hasil ini secara khusus menunjukkan bahwa strategi untuk mengatasi ketidakseimbangan kelas melalui pengambilan sampel yang berlebihan dan validasi yang cermat menggunakan *StratifiedKFold* dapat berdampak positif terhadap keakuratan model klasifikasi untuk teks yang terkait dengan kasus inses.

6. Evaluasi *F- measure* model: Tabel 9 menunjukkan hasil yang sangat baik, terutama untuk model dengan pemisahan data 80:20 dimana ukuran F1 mencapai 97%.

TABEL 9
PERBANDINGAN F1-SCORE MODEL KLASIFIKASI

Presentase pembagian	F1-Score
70:30	94%
80:20	97%

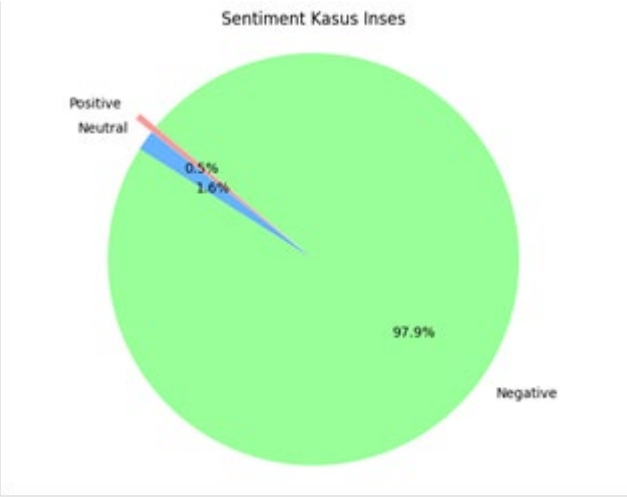
Hasil ini mengkofirmasi bahwa strategi pemrosesan ketidakseimbangan kelas dan validasi *StratifiedKFold* memiliki dampak positif pada akurasi model, sekaligus mencapai keseimbangan antara presisi dan perolehan dalam klasifikasi teks terkait kasus inses.

7. *Wordcould*: penelitian ini juga fokus pada visualisasi kata-kata yang biasa digunakan dalam *tweet* tentang kasus inses. Korban inses mempunyai pengalaman beragam, dengan pilihan kata untuk menggambarkan pengalaman tersebut memainkan peran penting dalam analisis ini. Pada *wordcould* Gambar 5, kata-kata tersebut seperti ”inses’ "anak", "insiden", "bayi", "konsekuensi", "rahim", "hubungan", "pembunuhan", "pemerkosaan".

Gambar 5. *wordcould* kasus inses

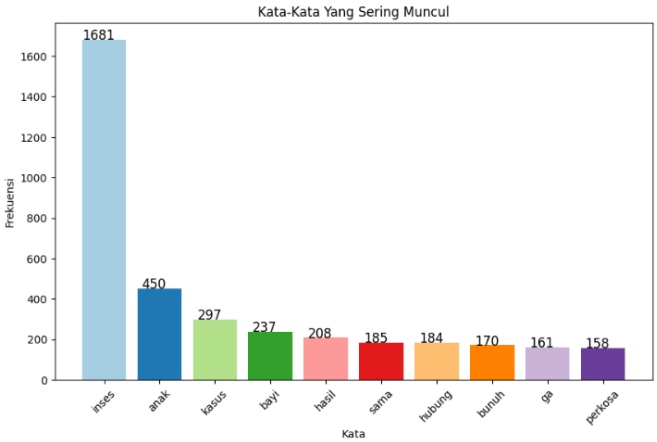
Pada Gambar 5 dapat dilihat bahwa kata yang sering muncul yaitu inses dengan kemunculan kata sebanyak 1681 kali, di ikuti dengan kata anak sebanyak 450 kata, kasus 297 kata, bayi sebanyak 237 kata.

8. Visualisasi Pie Chart: Visualisasi pie chart dibawah ini menggambarkan hasil analisis sentiment terkait kasus inses yang sedang dibahas di media sosial. Dalam chart ini, terdapat tiga kategori sentimen yang direpsentasikan oleh warna yang berbeda yaitu warna hijau sebagai repsentasi sentimen negatif, warna biru sebagai repsentasi sentimen netral, dan warna merah muda sebagai repsentasi sentimen positif. Dari total pie chart, mayoritas sentimen yang tercatat adalah negatif, mencapai 97,7%. Sentiment netral hanya mencakup 1,6% dari total chart, sementara sentiment positif hanya sebesar 0,5%. Dapat dilihat pada Gambar 6 dibawah ini.



Gambar 6. *Pie Chart* sentimen inses

Pada Gambar 7 menunjukkan bahwa mayoritas yang muncul di media sosial terkait kasus inses adalah negatif, menunjukkan adanya ketidaksetujuan, kecaman, atau perasaan negative terhadap kasus tersebut.



Gambar 7. Frekuensi kata yang sering muncul

IV. SIMPULAN

Berdasarkan hasil penelitian yang telah dilaksanakan, dapat disimpulkan bahwa tanggapan masyarakat terhadap kasus inses di Indonesia lebih menunjukkan sikap negative terhadap kasus inses, seperti yang ditunjukkan oleh analisis sentiment menggunakan *Naïve Bayes Classifier* dengan tingkat akurasi 94%. Strategi *oversampling* dan penerapan *StratifiedKFold* dalam *cross-validation* memainkan peran penting dalam pencapaian ini. Namun, algoritma *naïve bayes* masih memiliki kelemahan. Salah satunya adalah sensitivitasnya terhadap asumsi kemandirian fitur dan kemampuan untuk memahami konteks kalimat yang kompleks. Oleh karena itu, langkah-langkah untuk pengembangan selanjutnya harus mencakup penelitian lebih lanjut dalam konteks bahasa Indonesia dengan dataset yang lebih besar. Untuk meningkatkan keakuratan analisis sentiment dan pemahaman konteks, disarankan untuk menggabungkan dengan teknik lain seperti *Deep Learning*. Selain itu, evaluasi konten multimedia dan pengembangan dataset yang lebih baik dari berbagai sumber dapat menjadi langkah penting untuk merespons dinamika tren di media sosial dengan lebih baik. Dengan meningkatkan kesadaran publik tentang dampak negatif terkait kasus inses dan memberikan pendidikan seksual yang lebih baik, penulis dapat mengurangi insiden inses dan memperkuat pemahaman tentang batasan yang harus dijaga dalam hubungan keluarga. Dengan isu inses ini penelitian menyoroti peran media sosial dalam bentuk opini publik. Dengan memahami sentimen yang muncul di media sosial, masyarakat dapat merespons isu sensitif seperti inses dengan lebih baik, dan pengawasan konten di media sosial perlu diperketat untuk mengurangi penyebaran informasi yang merugikan.

DAFTAR PUSTAKA

[1] M. F. Fahrezi and A. A. Permana, "Sentimen Analisis Opini Masyarakat Pada Sosial Media Twitter Terhadap Organisasi Aksi Cepat Tanggap Menggunakan Naïve Bayes Classifier," vol. 11, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index
[2] S. O. Raida, N. Fathin, and Y. Pagayo, "PELAYANAN UPTD PEMBERDAYAAN PEREMPUAN DAN PERLINDUNGAN ANAK TERHADAP KASUS KEKERASAN SEKSUAL (INSES) PADA ANAK DI PROVINSI LAMPUNG," 2023.
[3] N. Hendrastuty, A. Rahman Isnain, and A. Yanti Rahmadhani, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," vol. 6, no. 3, 2021, [Online]. Available: http://situs.com
[4] D. Alita and A. Rahman, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," 2020.
[5] I. Sholekha, A. Faqih, and A. Bahtiar, "Sentiment Analysis of Public Opinion Covid-19 Vaccine Using Naïve Bayes and Random Forest Methods," JURNAL TEKNIK INFORMATIKA, vol. 15, no. 1, pp. 34–43, Jun. 2022, doi: 10.15408/jti.v15i1.24847.
[6] D. Alita, Y. Fernando, and H. Sulistiani, "IMPLEMENTASI ALGORITMA MULTICLASS SVM PADA OPINI PUBLIK BERBAHASA INDONESIA DI TWITTER," Jurnal TEKNOKOMPAK, vol. 14, no. 2, p. 86, 2020.
[7] E. P. Harahap, H. D. Purnomo, A. Iriani, I. Sembiring, and T. Nurtino, "Trends in sentiment of Twitter users towards Indonesian tourism: analysis with the k-nearest neighbor method," Computer Science and Information Technologies, vol. 5, no. 1, pp. 13–22, Mar. 2024, doi: 10.11591/csit.v5i1.pp13-22.
[8] R. Syaputra, R. Andryani, and D. Erlansyah, "Naïve Bayes Method for Text-Based Sentiment Analysis on Social Media," 2023.
[9] Z. Firmansyah and N. F. Puspitasari, "ANALISIS SENTIMEN MASYARAKAT TERHADAP VAKSINASI COVID-19 BERDASARKAN OPINI PADA TWITTER MENGGUNAKAN ALGORITMA NAIVE BAYES," Jurnal Teknik Informatika, vol. 14, no. 2, 2021, doi: 10.15408/jti.v14i2.24024.
[10] D. W. Ardras and A. Voutama, "ANALISIS SENTIMEN ANTI LGBT DI INDONESIA MELALUI MEDIA SOSIAL TWITTER," Jurnal Teknika, vol. 15, no. 1, pp. 23–28, Mar. 2023, doi: 10.30736/jt.v15i1.926.
[11] J. Teknika, R. K. Septiani, S. Anggraeni, and S. D. Saraswati, "Teknika 16 (02): 245-254 Klasifikasi Sentimen Terhadap Ibu Kota Nusantara (IKN) pada Media Sosial Menggunakan Naive Bayes," IJCCS, vol. x, No.x, pp. 1–5.
[12] A. Perdana, A. Hermawan, and D. Avianto, "Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan Naive Bayes Clasifier," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 11, no. 2, pp. 195–200, Jul. 2022, doi: 10.32736/sisfokom.v11i2.1412.
[13] J. V Sistem Komputer dan Kecerdasan Buatan Volume Nomor, R. Dwi Septiana, and A. Budi Susanto, "Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic Dan Particle Swarm Optimization," 2021.
[14] D. Darwis, E. Shintya Pratiwi, A. Ferico, and O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," 2020.
[15] P. Kumala Sari and R. Randy Suryono, "KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK ANALISIS SENTIMEN METAVERSE," 2024.