JOISIE (Journal Of Information Systems And Informatics Engineering)
Vol. 9, No.2, Desember 2025, Hlm 380-391
Received: 9 September 2025 Revised: 20 November 2025, Accepted: 16 Desember 2025
p- ISSN: 2503-5304
e- ISSN: 2527-3116

ANALISIS SENTIMEN DAN TOPIK KEKERASAN SEKSUAL PEREMPUAN DI X 2022-2025 BERBASIS *NAÏVE BAYES*-LDA

Ira Anggraini Siregar[1)], Fesa Asy Syifa Nurul Haq [2)]
[1,2]Prodi PJJ Sistem Informasi, Universitas Siber Asia, Jalan R.M Harsono No.1 RT09/04 Ragunan, Pasar Minggu, Jakarta Selatan, Jakarta, Indonesia
email: iraanggraini.siregarr@gmail.com

Abstract

Sexual violence against women remains a significant social issue in Indonesia. In recent years, X has become one of the platforms where the public actively expresses opinions, shares experiences, and discusses this topic. This study aims to analyze public sentiment and identify dominant themes in conversations related to sexual violence against women in Indonesia from 2020 to 2025. A total of 12,760 tweets were collected through crawling and processed using standard text-preprocessing techniques. Sentiment analysis was performed using the Naïve Bayes algorithm, while topic extraction was carried out using Latent Dirichlet Allocation (LDA). The results show that negative sentiment dominates with 75.8%, followed by positive sentiment at 15.9% and neutral sentiment at 8.3%. LDA reveals five major topics representing public reactions, legal issues, personal experiences, and discussions on various forms of violence. These findings offer insights into public perception and may contribute to the development of more responsive policies supporting survivors of sexual violence.

Keywords: Sentiment Analysis, Sexual Violence, Naïve Bayes, LDA, NLP

Abstrak

Kekerasan seksual terhadap perempuan masih menjadi isu sosial yang mendapat perhatian serius di Indonesia. Dalam beberapa tahun terakhir, X menjadi salah satu platform tempat masyarakat menyampaikan opini, berbagi pengalaman, dan berdiskusi mengenai isu ini. Penelitian ini bertujuan menganalisis sentimen publik serta mengidentifikasi topik utama dalam percakapan terkait kekerasan seksual terhadap perempuan di Indonesia selama periode 2020 hingga 2025. Sebanyak 12.760 tweet dikumpulkan melalui proses crawling dan diproses menggunakan teknik pra-pemrosesan teks. Analisis sentimen dilakukan dengan algoritma Naïve Bayes, sedangkan ekstraksi topik menggunakan metode Latent Dirichlet Allocation (LDA). Hasil menunjukkan bahwa sentimen negatif mendominasi dengan proporsi 75,8%, disusul sentimen positif sebesar 15,9% dan sentimen netral sebesar 8,3%. LDA menghasilkan lima topik utama yang mencerminkan reaksi publik, isu hukum, pengalaman pribadi, serta berbagai bentuk kekerasan yang dibahas warganet. Temuan ini memberikan gambaran mengenai persepsi publik dan dapat menjadi dasar dalam penyusunan kebijakan yang lebih responsif terhadap perlindungan perempuan.

Kata kunci: Analisis Sentimen, Kekerasan Seksual, Naïve Bayes, LDA, NLP

1. PENDAHULUAN

Indonesia merupakan salah satu negara berkembang dengan jumlah penduduk mencapai 284,4 juta jiwa yang terdiri dari 143,548 juta pria dan 140,890 juta perempuan. Dengan tingginya jumlah penduduk yang ada, membuat Indonesia menghadapi banyak tantangan sosial yang rumit, salah satunya adalah ketidaksetaraan *gender*, yang sering menempatkan perempuan pada posisi lebih rendah dari laki-laki dan memperlakukannya secara tidak adil dalam kehidupan sosial. Hal ini dibuktikan dengan hasil penelitian yang menunjukkan bahwa sistem patriarki masih banyak ditemukan dalam kehidupan sosial, dan menerima banyak penolakan dari masyarakat (Kurmasih et al., 2024).

Bukti nyata dari ketidaksetaraan ini tergambar jelas dengan maraknya kekerasan yang terjadi terhadap perempuan, terutama kekerasan seksual yang terus berlangsung hingga kini dan sering kali tidak mendapat tanggapan serta penanganan yang adil. Penanganan yang masih lemah, menjadi salah satu faktor hal ini terus terjadi.

https://doi.org/10.35145/joisie.v9i2.5528

Kekerasan seksual merupakan tindakan terlarang yang dilakukan tanpa persetujuan korban; secara paksa, mengancam dan intimidatif yang melibatkan bagian tubuh sensitif. Tindakan ini biasanya meninggalkan cedera fisik atau trauma kepada yang mengalaminya. Korban kekerasan seksual sebagian besar adalah perempuan, dan hal ini menunjukkan bahwa sistem perlindungan terhadap perempuan masih lemah.

Komisi Nasional Anti Kekerasan terhadap Perempuan (Komnas Perempuan), mencatat kasus kekerasan terhadap perempuan terus bertambah dalam lima tahun terakhir. Jumlahnya tercatat sebanyak 226,062 kasus pada 2020, kemudian meningkat menjadi 338,496 kasus pada tahun 2021, dan sebanyak 339,782 kasus pada tahun 2022. Tahun 2023 jumlah kasus sempat turun menjadi 289,111, namun kembali meningkat pesat pada tahun 2024 menjadi 445,502 kasus. Mayoritas kasus yang tercatat merupakan kekerasan seksual (Komnas Perempuan, 2024).

Sementara itu, belum ada laporan resmi tentang jumlah kekerasan seksual yang dialami perempuan pada tahun 2025. Namun sudah banyak berita *online* maupun *offline* yang melaporkan telah terjadi beberapa kasus kekerasan seksual terhadap perempuan dengan pelaku dari berbagai latar belakang. Lambat laun, tidak ada lagi tempat yang aman untuk perempuan. Pelaku kekerasan seksual sudah menjamur di semua lini kehidupan masyarakat.

Pemerintah sebenarnya telah menerbitkan berbagai regulasi maupun kebijakan untuk menindak pelaku kekerasan seksual, seperti Pasal 281, 289, dan 290 dalam Kitab Undang-Undang Hukum Pidana (Database Peraturan KUHP, 2012) serta Undang-Undang No. 12 Tahun 2022 tentang Tindak Pidana Kekerasan Seksual (UU TPKS, 2022). Namun implementasinya sering dinilai belum memadai sehingga belum memberikan efek jera maupun perlindungan yang optimal untuk perempuan. Kondisi ini menunjukkan perlunya pendekatan yang lebih komprehensif, bukan hanya dari aspek hukum, tetapi juga dari aspek sosial dan persepsi publik.

Di era digital saat ini, media sosial menjadi ruang baru di mana publik dapat berbagi pengalaman serta menyuarakan pendapat tentang masalah sosial di Indonesia, termasuk kekerasan seksual. X (sebelumnya Twitter) menjadi salah satu *platform* yang populer dengan lebih dari 60 juta pengguna aktif di Indonesia, dan menjadi tempat menyuarakan pendapat, baik berupa dukungan atau kritikan.

Berbagai penelitian sebelumnya telah menggunakan X dalam analisis persepsi publik, salah satu contohnya adalah penelitian yang menunjukkan sebagian besar pengguna X mendukung pengesehan RUU PKS, walaupun masih terdapat sentimen negatif (Hamidi et al., 2021).

Meskipun beberapa penelitian sebelumnya telah menggunakan data X untuk menganalisis isu kekerasan seksual, mayoritas hanya berfokus pada analisis sentimen tanpa mengeksplorasi struktur topik pembahasan yang ada di dalamnya. Selain itu, penelitian dengan rentang waktu panjang dan *dataset* besar masih terbatas sehingga belum mampu menggambarkan perubahan persepsi publik secara menyeluruh pada periode 2020-2025. Padahal metode *topic modeling* seperti *Latent Dirichlet Allocation* (LDA) terbukti efektif dalam mengidentifikasi pola topik tersembunyi dalam teks tidak terstruktur (Sahria et al., 2021).

Untuk mengisi celah tersebut, penelitian ini menggabungkan dua metode yaitu analisis sentimen menggunakan *Naïve Bayes* dan pemodelan topik menggunakan LDA dalam satu kerangka analisis yang akan lebih representatif dan komprehensif.

Kolaborasi dari kedua metode ini dalam konteks kekerasan seksual terhadap perempuan masih jarang dilakukan, sehingga penelitian ini dapat mengisi celah tersebut dengan menghadirkan analisis yang lebih menyeluruh.

Tujuan akhir dari penelitian yang menggabungan dua metode ini adalah dapat memberikan gambaran yang lebih jelas tentang persepsi publik terhadap isu kekerasan seksual yang terjadi kepada perempuan, memberikan wawasan yang dapat dijadikan dasar dalam penyusunan kebijakan yang lebih efektif, serta dapat berkontribusi dalam pengembangan sistem pemantauan isu sosial berbasis media sosial secara *real-time*.

2. METODE PENELITIAN

Penelitian ini menggunakan metode Kuantitatif dengan pendekatan *Data Mining* dan *Natural Language Processing* (NPL). Dalam proses penelitian digunakan dua teknik utama, yaitu penggunakan algoritma *Naïve Bayes* yang berperan untuk melakukan analisis sentimen masyarakat

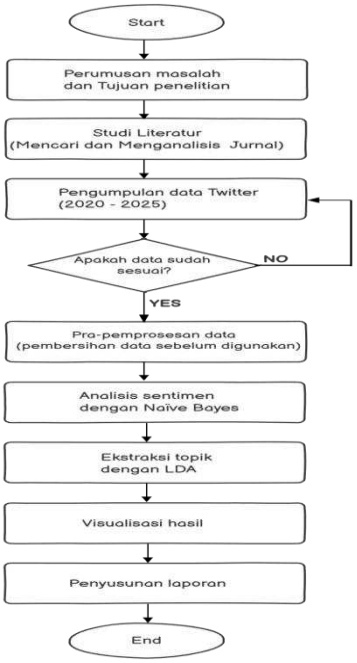
https://doi.org/10.35145/joisie.v9i2.5528

melalui percakapan yang diunggah di X. Algoritma ini akan menganalisis apakah sebuah tulisan mengandung makna negatif, positif, atau netral. Kemudian tiga sentimen ini akan dipresentasikan dalam sebuah grafik persentase.

Teknik kedua yang digunakan adalah metode LDA, yang berperan untuk menemukan topik-topik pembahasan yang sering muncul dalam sebuah percakapan di X. Sederhananya, LDA akan mencoba membaca semua data percakapan, lalu mengelompokkan kata-kata yang sering muncul secara bersama, menjadi beberapa topik. Hasil dari topik tesebut akan disajikan secara rapi dan mudah dipahami.

Pengumpulan data dilakukan hanya pada *tweet* yang bersifat publik. Indentitas pengguna seperti *username*, dan metadata pribadi dihapus pada tahapan pra-pemrosesan data. Penelitian ini mengikuti ketentuan pengguna data publik berdasarkan kebijakan X untuk kepentingan akademik.

Berikut ini adalah diagram alur yang menggambarkan langkah-langkah yang akan dilakukan selama penelitian (Gambar 1.).



Gambar 1. Tahapan Penelitian

2.1 STUDI LITERATUR

Studi literatur dilakukan dengan membaca jurnal-jurnal yang berkaitan dengan topik penelitian, dan kemudian membuat ringkasan terkait isi dari jurnal, metode yang digunakan, hingga kesimpulan yang diperoleh. Setelah ringkasan dibuat, proses berikutnya adalah menentukan *research gap* yang ditemukan dari kumpulan jurnal tersebut. *Research gap* atau kekosongan penelitian ini membantu dalam memperkuat alasan penelitian ini mengangkat isu kekerasan seksua terhadap perempuan dengan menggunakan metode *Naïve Bayes* dan LDA, karena belum ditemukan penelitian yang menggabungkan dua metode tersebut.

2.2 PENGUMPULAN DATA X

Pengumpulan data dilakukan secara bertahap, yang dibagi per tahun. Proses pengumpulan data menggunakan *Tweet-Harvest* yang merupakan salah satu skrip Python yang sering digunakan untuk melakukan *scraping tweet* berdasarkan kata kunci tertentu, dengan bantuan *Twitter Auth Token* untuk akses konten X. *Tweet-Harvest* bekerja dengan cara menampilkan *tweet* terbaru terlebih dahulu. Hal ini yang mendasari *crawling* data dilakukan bertahap.

Pengumpulan data yang dilkaukan sekaligus dengan rentang waktu yang panjang yaitu dari tahun 2020 hingga 2025, akan menghasilkan data yang didominasi oleh data terbaru yaitu tahun 2025. Agar distribusi data yang dihasilkan lebih merata, maka proses *crawling* dibagi per tahun.

https://doi.org/10.35145/joisie.v9i2.5528

Selain itu, X juga memberikan batasan jumlah *tweet* yang dapat diambil dalam satu waktu, sehingga proses pengumpulan data pada dasarnya tidak biisa dilakukan sekaligus.

Proses ini, digunakan banyak kata kunci yang relevan dengan isu kekeran seksual terhadap perempuan di Indonesia, beberapa diantaranya "pelecehan seksual", "kekerasan seksual", "pemerkosaan", "pelecehan terhadap perempuan", "pelecehan online", "kekerasan pada perempuan tahun ini", dan berbagai tagar seperti:
"#StopKekerasanSeksualPerempuan", "#KeadilanUntukPerempuan", "#PerempuanMerdeka", "PerempuanBukanObjek".
Data-data yang berhasil dikumpulkan per tahun, disimpan di Google Drive menjadi beberapa dokumen, yang kemudian dokumen-dokumen itu digabungankan menjadi satu *dataset* utuh yang diurutkan secara *ascending* agar dapat merepresentasikan data mentah dari rentah waktu 2020-2025. Pada proses ini berhasil dikumpulkan sebanyak **13,518** *tweet* sebagai data mentah, yang akan digunakan untuk proses penelitian.

2.3 PRA-PEMROSESAN DATA

Proses ini bertujuan untuk memastikan data sudah bersih dan siap digunakan untuk analisis. Terdapat beberapa tahapan pada proses ini, diantaranya:

a. *Case folding*

Proses ini bertujuan untuk mengubah semua teks menjadi huruf kecil. Hal ini dilakukan untuk menghindari perbedaan makna antara kata yang sama namun memiliki kapitalisasi berbeda, misalnya kata "Kekerasan" dan "kekerasan". Selanjutnya dilakukan juga pembersihan tambahan berupa penghapusan spasi yang tidak perlu di awal dan akhir kalimat, serta menghapus semua spasi yang berlebih.

b. Menghapus Karakter

Tahapan selanjutnya pada proses pembersihan data yaitu menghapus karakter yang tidak relevan, yang dapat mengganggu proses analisis. Proses ini meliputi menghapus tautan atau URL yang terdapat dalam teks, menghapus *mention* atau *@username*, menghapus angka, tanda baca, serta karakter non-alfabet selain spasi. Penghapusan karakter ini dilakukan untuk menyederhanakan teks menjadi rangkaian kata saja. Karena dalam banyak kasus analisis teks, karakter ini tidak memberikan informasi yang bermanfaat untuk proses analisis.

c. Cek dan Hapus Data Duplikat

Proses ini bertujuan untuk melihat apakah terdapat data yang sama pada baris data. Jika ditemukan data duplikat pada beberapa baris data, maka data tersebut akan dihapus. Hal ini bertujuan untuk menghindari adanya data yang sama pada *dataset*. Pada proses ini terdapat 785 data duplikat dan 12.760 data normal. Kumpulan data normal sebanyak **12.760** ini merupakan jumlah *dataset* akhir yang akan digunakan selama proses penelitian berlangsung.

d. Penerjemahan Data

Proses ini bertujuan untuk melihat apakah terdapat data yang mengandung bahasa inggris, dan menerjemahkannya ke bahasa Indonesia. Proses ini membantu untuk memperoleh keselarasan bahasa dalam *dataset*. Pada proses ini terdapat 21 data yang mengandung bahasa inggris.

e. *Stopword Removal*

Proses ini bertujuan untuk menghapus kata-kata yang dianggap tidak punya makna penting dalam analisis teks, yang biasanya merupakan kata umum seperti "yang", "dengan", "dan", "pada", "saya", "di", "untuk" dan sejenisnya. Kata umum seperti ini dapat mengganggu proses analisis sentimen dan topik karena tidak memberikan kontribusi informasi yang signifikan.

f. *Tokenisasi*

Tokenisasi berperan untuk memecah kalimat menjadi unit-unit kecil yang disebut *token* atau kata per kata. Proses ini sangat penting karena NLP hanya bekerja secara efektif jika inputnya sudah berbentuk token.

g. *Stemming*

Tokenisasi Stemming berperan untuk mengubah kata ke bentuk dasarnya. Proses ini membantuk mengurangi variasi kata dan membuat analisis teks menjadi lebih konsisten.

h. *Detokenisasi*

https://doi.org/10.35145/joisie.v9i2.5528

Proses *detokenisasi* ini berperan untuk menggabungkan kembali token-token pada proses sebelumnya menjadi sebuah kalimat atau teks utuh. Proses ini bertujuan agar data kembali ke bentuk semula, berupa kalimat utuh, namun hanya menyisakan kata dasar yang dapat digunakan pada proses selanjutnya.

2.4 ANALISIS SENTIMEN DENGAN *NAÏVE BAYES*

Setelah tahapan pra-pemrosesan data dilalukan dan dipreroleh sebanyak 12.760 *dataset* yang sudah bersih. Proses selanjutnya yaitu analisis sentimen publik menggunakan *Naïve* Bayes, yang dibagi menjadi beberapa tahapan berikut.

a. *Generate* Data Sampel

Sebelum dilakukan analisis sentimen pada seluruh *dataset*, dibutuhkan data latih atau data sampel yang dilabeli secara manual, yang nantinya data latih ini digunakan sebagai bahan untuk melatih model sebelum digunakan untuk memprediksi sentimen pada keseluruhan *dataset*.

Pengambilan data latih dihasilkan secara acak sebanyak 1000 data *tweet* dari *dataset* yang sudah melalui tahapan pra-pemrosesan data.

Data sampel ini disimpan dalam sebuah file .csv di Google Drive dan dilabeli secara manual ke dalam tiga kategori sentimen, yaitu **positif**, **netral**, dan **negatif**, dan menghasilkan proporsi data seperti pada Table 1.

Tabel 1. Perbandingan Sentimen Data Latih

Positif	Netral	Negatif
69	45	886

b. *Training* Data Sampel

Setelah proses pelabelan manual dilakukan pada 1000 data sampel, 80% dari data digunakan untuk proses selanjutnya yaitu melatih data sampel menggunakan *Naïve Bayes* atau disebut dengan pelatihan model analisis sentimen. Sebelum model dilatih, terlebih dahulu dilakukan teknik *oversampling*, yaitu menggandakan data dari sentimen minoritas (**positif dan netral**), hingga jumlahnya setara dengan kelas mayoritas. Tujuannya agar model dilatih dengan data yang seimbang, sehingga tidak bias pada salah satu sentimen. Ketidakseimbangan data antar sentimen dapat menyebabkan model cenderung memprediksi semua data ke kelas yang dominan, yaitu sentimen negatif. Berikut ini tabel proporsi data setelah dilakukan proses *balancing*.

Tabel 2. Proporsi Setelah *Balancing*

Positif	Netral	Negatif
886	886	886

Setelah data seimbang, teks pada *tweet* diubah ke bentuk numerik menggunakan metode TF-IDF (*Term Frequency - Inverse Document Frequency*). Metode ini bekerja dengan memberikan bobot lebih tinggi pada kata-kata yang dianggap penting dan sering muncul dalam satu *tweet*, namun jarang ditemukan di *tweet* lain. Sebagai contoh, kata "pemerkosaan" muncul lebih sedikit di *tweet*, dianggap lebih bermakna dibanding kata "kekerasan" yang muncul hampir di semua *tweet*. Dengan cara ini, model dapat lebih fokus pada kata-kata yang benar-benar khas dan relevan terhadap sentimen, bukan sekadar kata umum.

Model *Naïve Bayes* kemudian dilatih menggunakan data yang sudah berbentuk numerik. Cara kerjanya adalah model mempelajari pola hubungan antara kata-kata dalam teks dan label sentimennya, agar mampu mengenali sentimen pada data baru. Setelah pelatihan selesai, model beserta TF-IDF *vectorizer* disimpan ke dalam file agar dapat digunakan kembali dalam proses prediksi tanpa perlu melatih ulang dari awal.

c. Prediksi Sentimen Seluruhan Data

Proses selanjutnya yaitu melakukan prediksi sentimen terhadap seluruh *dataset* yang berjumlah **12.760** *tweet*. Model *Naïve Bayes* dan TF-IDF *vectorizer* yang sebelumnya telah disimpan, dimuat kembali untuk memproses *dataset*. Teks pada setiap *tweet* diubah menjadi numerik menggunakan TF-IDF, dan kemudian sentimen dari setiap baris data diprediksi oleh

https://doi.org/10.35145/joisie.v9i2.5528

model berdasarkan pola yang telah dipelajari selama pelatihan. Kemudian hasil prediksi ditambahkan ke *dataset* dan menghasilkan proporsi sentimen seperti pada Tabel 3. berikut ini.

Tabel 3. Hasil Sentimen Analisis

Positif	Netral	Negatif
2027	1056	9677

d. Evaluasi Model

Setelah model *Naïve Bayes* dilatih menggunakan data berlabel, dilakukan proses evaluasi untuk mengukur seberapa baik model dalam memprediksi sentimen dari total 1.000 data, sebanyak 20% digunakan sebagai data uji. Evaluasi dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *f1-score* yang dihitung otomatis menggunakan *library* sklearn. Secara teknik, metrik dihitung berdasarkan rumus:

Akurasi berfungsi untuk mengukur proporsi prediksi yang benar dari seluruh data uji.

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision berfungsi untuk menunjukkan ketepatan model dalam memprediksi suatu sentimen.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall berfungsi untuk menunjukkan seberapa banyak data sebenarnya yang berhasil dikenali model.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score merupakan kombinasi *precision* dan *recall* yang berguna saat data tidak seimbang.

$$F1 = 2\ X\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

2.5 EKSTRAKSI TOPIK DENGAN LDA

Untuk menemukan topik utama dalam *tweet*, digunakan metode *Latent Dirichlet Allocation* (LDA). Pada penelitian ini, LDA tidak bekerja berdasarkan tahun, waktu *tweet*, pengguna, atau lokasi, karena metode ini memang fokus pada isi teks, karena tujuannya adalah menemukan tema yang paling sering muncul, bukan menganalisis tren berdasarkan waktu atau metadata lain.

LDA bekerja dengan cara mencari kata-kata yang sering muncul bersama dalam *tweet*, dan mengelompokkannya menjadi topik. Data yang digunakan adalah hasil *stemming* yang sudah dikonversi menjadi list kata.

Dari data tersebut dibentuk *dictionary* (kumpulan kata unik) dan *corpus* (representasi numerik dari teks), yang diperlukan untuk proses pelatihan model.

Pada tahapan ini, terdapat 5 topik utama yang akan ditemukan dengan bantuan LDA.

Pemilihan jumlah topik sebanyak 5 dilakukan melalui proses eksplorasi awal dengan mencoba beberapa alternatif jumlah topik yang umum digunakan dalam penelitian LDA. Peneliti mengevaluasi hasil keluaran model secara manual dengan membandingkan kejelasan kata-kata dominan dan tingkat tumpang tindih antartopik. Pada pengujian 3-4 topik, kelompok kata yang muncul terlalu luas sehingga beberapa isu penting bergabung. Sebaliknya, ketika jumlah topik ditambah, beberapa topik justri terlihat terlalu mirip dan tidak memberikan informasi tambahan yang berarti.

Pada akhirnya 5 topik dipilih karena menghasilkan pemisahan tema paling logis dan koheren dari segi makna, dan sesuai dengan keragaman isu kekerasan seksual yang muncul dalam *dataset*.

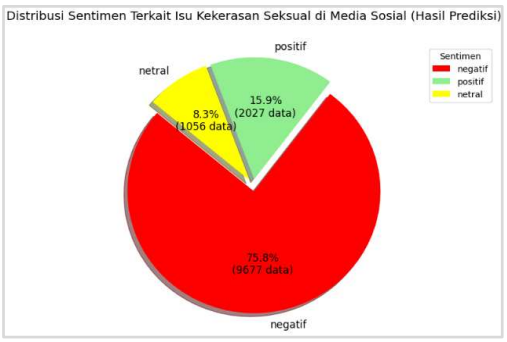
https://doi.org/10.35145/joisie.v9i2.5528

3. HASIL DAN PEMBAHASAN

Berikut ini hasil dari analisis sentimen dan ekstraksi topik yang dihasilkan dalam penelitian.

3.1 PROPORSI DAN DISTRIBUSI SENTIMEN

Pada proses analisis sentimen publik menggunakan *Naïve Bayes*, diperoleh jumlah sentimen positif sebanyak **2027** data (**15.9%**), dan **1056** data (**8.3%**) sentimen netral, dan sebanyak **9677** data (**75.8%**) memiliki sentimen negatif. Proporsi tersebut divisualisasikan dalam bentuk diagram pada Gambar 2.



Gambar 2. Proporsi dan Distribusi Sentimen

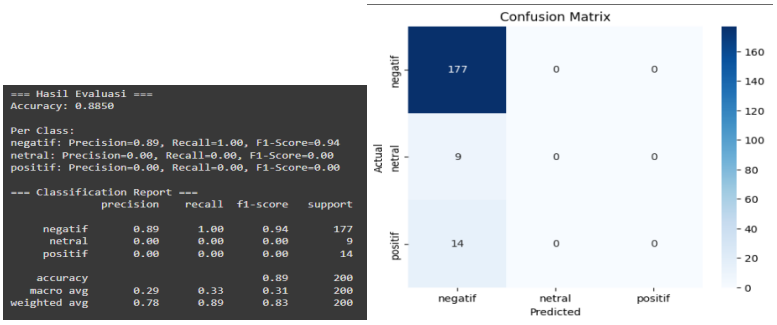
Dari Gambar 2 di atas, dapat disimpulkan bahwa lebih dari 50% percakapan publik di X didominasi oleh **sentimen negatif**, yang mengecam dan menganggap bahwa kekerasan seksual terhadap perempuan adalah tindakan yang sangat buruk dan tidak bisa dinormalisasi. Hal ini sangat wajar, karena isu ini merupakan isu yang sensitif, menyangkut ketidakadilan dan kemarahan publik.

Sementara itu **sentimen positif**, umumnya berisi tentang dukungan publik terhadap korban, apresiasi atas upaya penyelesaian kasus, dan gerakan positif yang dibuat untuk meningkatkan kesadaran masyarakat terhadap isu ini. **Sentimen netral**, biasanya berupa berita atau informasi tanpa ekspresi opini yang kuat, serta pertanyaan publik yang masih ambigu terhadap isu kekerasan seksual.

Hasil penelitian yang menunjukkan bahwa sentimen negatif mendominasi percakapan publik terkait kekerasan seksual terhadap Perempuan di X. Dominasi ini sejalan dengan temuan Kurmasih et al. (2024) dan Hamidi et al. (2022), yang menyatakan bahwa isu kekerasan seksual berbasis *gender* memicu reaksi keras dan emosional dari publik, terutama karena lemahnya penegakan hukum dan tingginya rasa ketidakadilan. Selain itu, kegagalan model mengenali sentimen positif dan netral pada evaluasi awal, konsisten dengan teori *class imbalance* (Google Developers, 2025), di mana distribusi data yang tidak seimbang membuat model lebih cenderung mengklasifikasikan seluruh teks ke kelas dominan, yaitu sentimen negatif.

3.2 HASIL EVALUASI MODEL

Berikut ini gambaran hasil evaluasi model yang diperoleh seperti Gambar 3.



Gambar 3. Hasil Evaluasi Model

Hasil evaluasi menunjukkan bahwa model *Naïve Bayes* memiliki akurasi **88,5%**. Namun performanya tidak merata antar kelas. *Precision* dan recall tertinggi hanya muncul pada sentimen negatif (*precesion* 0.89 dan *recall* 1.00), sedangkan sentimen positif dan netral sama sekali tidak

https://doi.org/10.35145/joisie.v9i2.5528

berhasil dikenali. Seluruh *tweet* positif dan netral diprediksi sebagai negatif, yang terlihat jelas pada *confusion matrix* di mana hanya kelas negative yang terisi.

Kondisi seperti ini disebabkan oleh ketidakseimbangan distribusi data latih yang sangat ekstrem, yakni 886 data negatif dibandingkan dengan data positif yang hanya 69 data dan 45 data netral. Ketimpangan ini membuat model mengalami *majority class* bias, yaitu kecenderungan memprediksi kelas terbanyak sebagai jawaban untuk hampir semua data.

Selain itu, *tweet* positif dan netral pada isu kekerasan seksual biasanya menggunakan bahasa yang lebih halus, informatif, atau ambigu sehingga tidak menghasilkan pola yang kuat untuk dibedakan, berbeda dengan *tweet* negatif yang umumnya lebih eksplisit secara emosional.

Evaluasi model tersebut dilakukan pada data berlabel asli tanpa proses *balancing* terlebih dahulu. Berikut ini adalah gambaran hasil evaluasi model pada 20% data sampel yang sebelumnya dilabeli secara manual.

Tabel 4. Hasil Sentimen Analisis

Aktual\Prediksi	Positif	Netral	Negatif
Positif	0	0	14
Netral	0	0	9
Negatif	0	0	177

Setelah dilakukan *oversampling* untuk menyeimbangkan jumlah data per kelas, model menjadi lebih stabil dan mampu memberikan hasil prediksi yang lebih proporsional. Model inilah yang kemudian digunakan untuk memprediksi keseluruhan *dataset* agar tidak bias terhadap salah satu sentimen.

3.3 VALIDASI SILANG

Validasi silang berfungsi untuk menguji performa model yang dibagi dalam lima *fold*. Hasil validasi silang menunjukkan bahwa akurasi model pada setiap *fold* secara berturut-turut adalah 0.89, 0.885, 0.885, 0.885, dan 0.885, dengan rata-rata akurasi sebesar **88,6%**. Nilai ini menunjukkan bahwa model memiliki performa yang stabil dan tidak hanya cocok pada data tertentu saja. Akurasi per *fold* digambarkan seperti yang terlihat pada Gambar 4 di bawah ini.

```
Akurasi per fold: [0.89  0.885 0.885 0.885 0.885]
Rata-rata akurasi: 0.8859999999999999
```

Gambar 4. Hasil Validasi Silang

3.4 VISUALISASI TOPIK DENGAN LDA

Berikut ini hasil ekstraksi topik menggunakan LDA yang divisualisasikan menggunakan *Word Cloud* seperti Gambar 5.

Gambar 5. Visualisasi Word Cloud

https://doi.org/10.35145/joisie.v9i2.5528

Gambar 5. menampilkan hasil visualisasi *word cloud* dari lima topik utama yang dihasilkan oleh model LDA berdasarkan kumpulan *tweet* terkait isu kekerasan seksual terhadap perempuan selama lima tahun terakhir. Setiap *word cloud* menunjukkan kata-kata yang sering muncul bersama dalam *tweet*, dan kata-kata tersebut membentuk interpretasi tertentu.

Interpretasi topik berikut ini merupakan hasil analisis pribadi peneliti dengan mengamati dan mengelompokkan makna kata-kata dominan yang muncul pada setiap topik. LDA sendiri hanya menghasilkan kumpulan kata per topik tanpa memberi nama atau label topik secara otomatis. Berikut ini rincian masing-masing topik:

a. Topik 1: kekerasan parah dan ekspresi marah

Word Cloud pada topik ini didominasi oleh kata "perkosa", "bunuh", "yg", dan "orang", dan juga kata-kata kasar dan kemerahan seperti "lu" dan "anjing", yang menggambarkan reaksi emosional yang tinggi dari publik terhadap kasus kekerasan seksual.

b. Topik 2: Perempuan sebagai korban kekerasan

Pada topik ini menonjolkan kata "keras", "perempuan", "leceh", dan "korban", yang mana dapat diartikan topik ini menggambarkan tentang perempuan yang menjadi korban kekerasan seksual. Kata "kampus", dan "lingkungan" juga muncul, yang menunjukkan bahwa kasus terjadi di lingkungan pendidikan.

c. Topik 3: penanganan hukum dan aparat

Topik ini menunjukkan kata "polisi", "duga", "kasus", dan "pidana", yang berfokus pada proses hukum, mulai dari pelaporan dan penanganan kasus oleh aparat. Kata seperti "anggota", "oknum", dan "tangkap", memperjelas bahwa pada kasus kekerasan seksual ini ada pihak berwenang yang diduga terlibat dan menjadi oknum.

d. Topik 4: opini publik serta pengalaman

Pada topik ini terdapat kata-kata yang muncul cukup dominan, seperti "yg", "seksual", "perkosa", dan "aja. Selain itu muncul juga kata "aku", "mau", dan "udah" yang menunjukkan bahwa beberapa *tweet* bersifat personal yang berisi bentuk opini pribadi, dan juga merepresentasikan pengalaman pribadi dan reaksi publik.

e. Topik 5: kekerasan seksual terhadap anak dan bentuk lain dari kekerasan seksual

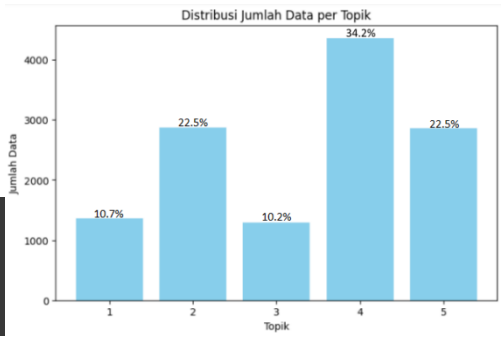
Topik kelima menunjukkan bentuk lain dari isu kekerasan seksual. Ini tergambar dari kata "anak", "fisik", "psikis", "pelajar", yang menunjukkan terdapat kekerasan dalam bentuk lain yaitu fisik dan mental, yang rentan terjadi kepada anak-anak. Ada juga kata "hukum", "tindak", dan "kasus" yang mendorong agar semua bentuk kekerasan bisa ditindak secara hukum.

Analisis topik menggunakan LDA di atas memperlihatkan lima pola utama yang mencerminkan dinamika percakapan publik, mulai dari luapan kemarahan, narasi korban, kritik hukum, pengalaman pribadi, hingga kekerasan terhadap anak pola ini sejalan dengan penelitian Sahria et al. (2020) yang menunjukkan bahwa LDA mampu mengungkap struktur diskusi yang berlapis dalam isu sosial. Keberadaan topik pengalaman pribadi juga mengindikasikan bahwa X berfungsi sebagai ruang berbagi dan advokasi informal bagi para pengguna, terutama dalam isu sensitive seperti kekerasan seksual.

3.5 VISUALISASI DISTRIBUSI TOPIK

Berikut ini Gambaran distribusi kemunculan dari setiap topik (Gambar 6).

```
Distribusi Jumlah Data per Topik:
Dominant_Topic
1    1367
2    2870
3    1298
4    4358
5    2867
```



Gambar 6. Visualisasi Distribusi Topik

https://doi.org/10.35145/joisie.v9i2.5528

Gambar 6 menunjukkan distribusi jumlah kemunculan pada setiap topik. Topik 4 merupakan topik yang paling dominan muncul sebanyak 4,358 kali atau sekitar 34,2% dari total *tweet*. Kemudian Topik 2 dan Topik 5 memiliki persentase kemunculan yang seimbang, yaitu 2.870 dan 2.867 data, atau 22.5%. Sedangkan Topik 1 dan Topik 3 jumlah kemunculannya sebesar 1.367 data (10.7%) dan 1.298 data (10,2%).

Distribusi ini memberikan Gambaran bahwa Sebagian besar pembahasan publik di media social berkaitan dengan Topik 4, yang sebelumnya telah dianalisis sebagai opini public serta pengalaman personal dalam menanggapi isu kekerasan seksual.

Hal ini menguatkan bahwa keterlibatan emosi dan pengalaman pribadi merupakan bentuk reaksi yang paling banyak disuarakan publik.

4. SIMPULAN

Berdasarkan hasil analisis sentimen terhadap percakapan publik mengenai kekerasan seksual terhadap perempuan di Indonesia periode Januari 2020-Juni 2025, ditemukan bahwa 75,8% pengguna X menunjukkan sentime negatif. Dominasi ini menggambarkan penolakan kuat masyarakat terhadap kekerasan seksual dan kekecewaan terhadap situasi sosial-hukum yang melingkupnya. Sentimen positif sebesar 15,9% terutama berisi dukungan kepada korban dan apresiasi terhadap gerakan edukasi, sedangkan 8,3% sentimen netral muncul pada unggahan yang bersifat informatif tanpa opini pribadi.

Model *Naïve Bayes* yang digunakan dalam penelitian ini mampu mengenali seluruh kategori sentimen secara lebih seimbang setelah penanganan ketidakseimbangan data, dengan akurasi akhir 89%. Hal ini menunjukkan bahwa model layak digunakan untuk prediksi sentimen pada isu serupa. Selain itu, pemodelan LDA berhasil mengungkap struktur diskusi publik melalui lima kelompok topik yang stabil dan relevan.

Penelitian ini memiliki beberapa keterbatasan. Pertama, data X tidak menyediakan informasi lokasi secara pasti sehingga analisis tidak dapat mengidentifikasi persebaran geografis sentimen. Kedua, proses *scraping* bergantung pada API pihak ketiga yang memiliki batasan kueri sehingga berpotensi membatasi jumlah dan variasi data. Ketiga, ketidakseimbangan data menyebabkan perlunya *oversampling*, yang dapat mempengaruhi generalisasi model. Keempat, penggunaan *Naïve Bayes* memiliki keterbatasan dalam menangani bahasa informal dan konteks ironi yang umum ditemukan pada media sosial X.

Penelitian selanjutnya dapat mengembangkan model berbasis transformer seperti BERT untuk meningkatkan akurasi konteks. *Dataset* yang lebih besar dan bersumber dari beberapa *platform* media sosial dapat memberikan gambaran yang lebih kaya. Analisis temporal juga dapat ditambahkan untuk melihat perubahan dinamika sentimen dari waktu ke waktu, serta eksplorasi teknik geolokasi berbasi profil atau konten untuk memahami persebaran opini publik secara lebih mendalam.

5. DAFTAR PUSTAKA

Ainiyah, K., & Holle, K. F. H. (2024). Analisis sentimen terhadap Permendikbud Ristek Nomor 30 Tahun 2021 pada media sosial Twitter menggunakan metode lexicon-based dan Multinomial Naïve Bayes. *Jurnal Ilmiah Informatika, 7*(1), 29–40. https://journal.ibrahimy.ac.id/index.php/JIMI/article/view/1589

Badan Pusat Statistik. (2025). *Jumlah penduduk menurut kelompok umur dan jenis kelamin*. https://www.bps.go.id/en/statisticstable/3/WVc0MGEyMXBkVFUxY25KeE9HdDZkbTQzWkVkb1p6MDkjMw==/jumlahpenduduk-menurut-kelompok-umur-dan-jenis-kelamin.html?year=2025

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons https://archive.org/details/cochran-1977-sampling-techniques

Dao, S. D., & Marian, R. (2011, March 16–18). *Optimisation of precedence-constrained production sequencing and scheduling using genetic algorithms*. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong.

Debora, U. R., Pratama, I. P. A., & Sasmita, G. M. A. (2024). Sentiment analysis of domestic violence issues on Twitter using Multinomial Naïve Bayes and Support Vector Machine. *JITTER:*

https://doi.org/10.35145/joisie.v9i2.5528

Jurnal Ilmiah Teknologi dan Komputer, 4(3), 1992–2000. https://ojs.unud.ac.id/index.php/jitter/article/view/110096/53135

Google Developers. (2025). *Classification: Accuracy, precision, recall, and related metrics*. Google Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Hamidi, G. D., Bestari, F. A., Situmorang, A., & Rakhmawati, N. A. (2022). Sentiment analysis on the ratification of Penghapusan Kekerasan Seksual Bill on Twitter. *JUTISI: Jurnal Teknik Informatika dan Sistem Informasi, 8*(3), 173–180. https://journal.maranatha.edu/index.php/jutisi/article/view/4051/2046

Komnas Perempuan. (2024). *CATAHU 2023: Peluang penguatan sistem penyikapan di tengah peningkatan kompleksitas kekerasan terhadap perempuan*. Komisi Nasional Anti Kekerasan terhadap Perempuan.

Kurmasih, M., Rahaningsih, N., Dana, R. D., & Nuris, N. D. (2024). Analisis sentimen terhadap opini patriarki menggunakan metode Naïve Bayes. *JATI: Jurnal Mahasiswa Teknik Informatika, 8*(3), 3095–3100. https://ejournal.itn.ac.id/jati/article/view/8798/5480

Republik Indonesia. (n.d.). *Kitab Undang-Undang Hukum Pidana (KUHP)*. https://peraturan.bpk.go.id

Republik Indonesia. (2022). *Undang-Undang Nomor 12 Tahun 2022 tentang Tindak Pidana Kekerasan Seksual (UU TPKS)*. https://peraturan.bpk.go.id/Home/Details/204576/uu-no-12-tahun-2022

Sahria, Y., & Fudholi, D. H. (2020). Analysis of health research topics in Indonesia using the LDA topic modelling method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 4*(2), 336–344. https://jurnal.iaii.or.id/index.php/RESTI/article/view/1821/236

Saemani, R. A., & Setiyawati, N. (2020). Sentiment analysis on the Permendikbud concern: Prevention and treatment of sexual violence in higher educational environments using Support Vector Machine (SVM). *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 8*(1), 65–71. https://ejournal.nusamandiri.ac.id/index.php/jitk/article/view/2807/995

Salsabilla, T., & Alita, D. (2024). Analisis sentimen inses di social media menggunakan algoritma Naïve Bayes. *Jurnal Informatika: Jurnal Pengembangan IT, 9*(3), 271–280. https://ejournal.poltekharber.ac.id/index.php/informatika/article/view/6611/3235

Simangunsong, S. H., Arandito, S., Daeng, M. F., Fauziyah, A., & Yunus, S. R. (2025, June 16). Gunung es kekerasan seksual dan tantangan proses hukum ramah perempuan. *Kompas.id*. https://www.kompas.id/artikel/gunung-es-kekerasan-seksual-dan-tantangan-proses-hukum-ramah-perempuan

Kemp, S. (2024, January 31). Digital 2024: Global digital report. We Are Social & Meltwater. https://wearesocial.com/id/blog/2024/01/digital-2024/

World Health Organization. (2024). *Sexual violence*. https://apps.who.int/violence-info/sexual-violence/

https://doi.org/10.35145/joisie.v9i2.5528