



Pemodelan Topik pada Komentar Media Sosial X menggunakan Latent Dirichlet Allocation

Ardelia Adzra, Safitri Juanita*

Fakultas Teknologi Informasi, Program studi Sistem Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: ¹2212500645@student.budiluhur.ac.id, ^{2,*}safitri.juanita@budiluhur.ac.id

Email Penulis Korespondensi: safitri.juanita@budiluhur.ac.id

Abstrak—Pelecehan seksual merupakan isu sosial yang banyak dibicarakan di media sosial X. Namun, tingginya volume komentar yang bersifat tidak terstruktur menyulitkan proses identifikasi topik pembahasan secara manual. Penelitian ini bertujuan mengidentifikasi topik utama pada komentar pelecehan seksual di media sosial X menggunakan metode Latent Dirichlet Allocation (LDA). Data yang digunakan berupa komentar mengenai topik pelecehan seksual yang dikumpulkan dari media sosial X pada periode 2024–2026. Tahapan penelitian meliputi pengumpulan data, persiapan data, pembentukan *dictionary* dan *corpus*, pemodelan LDA dengan penyetelan *hyperparameter tuning*, evaluasi menggunakan *coherence score*, serta interpretasi topik berdasarkan kata kunci dominan dan data representatif. Hasil penelitian menunjukkan bahwa model LDA terbaik terdiri dari empat topik dengan nilai *coherence score* sebesar 0,517. Keempat topik tersebut diinterpretasikan sebagai Penanganan Kasus Pelecehan Seksual di Lingkungan Pendidikan, Pengalaman dan Dampak Psikologis Korban, Kasus Pelecehan Seksual di Perguruan Tinggi, serta Perlindungan terkait Pelecehan Seksual. Temuan ini menunjukkan bahwa metode LDA mampu mengidentifikasi topik utama pada komentar pelecehan seksual serta membantu mengelompokkan data media sosial yang tidak terstruktur menjadi informasi yang lebih mudah dipahami. Kontribusi penelitian ini adalah menyajikan pendekatan pemodelan topik berbasis LDA dengan penyetelan *hyperparameter* untuk mengelompokkan komentar pelecehan seksual di media sosial X yang tidak terstruktur menjadi topik-topik utama yang lebih mudah dipahami. Hasil pemetaan topik ini dapat menjadi dasar dalam memahami isu yang paling banyak mendapat perhatian masyarakat, serta berpotensi dimanfaatkan untuk mendukung pengembangan layanan dukungan korban, termasuk pada sistem telemedis.

Kata Kunci: Analisis Teks; LDA; Media sosial X; Pelecehan Seksual; Pemodelan Topik

Abstract—Sexual harassment is a social issue widely discussed on the social media platform X. However, the high volume of unstructured comments makes it difficult to manually identify the main topics of discussion. This study aims to identify the main topics in comments related to sexual harassment on X using the Latent Dirichlet Allocation (LDA) method. The data used consist of comments on the topic of sexual harassment collected from X during the 2024–2026 period. The research stages include data collection, data preparation, dictionary and corpus construction, LDA modeling with hyperparameter tuning, evaluation using coherence score, and topic interpretation based on dominant keywords and representative data. The results show that the best LDA model consists of four topics with a coherence score of 0.517. These four topics are interpreted as Handling Cases of Sexual Harassment in Educational Environments, Victims' Experiences and Psychological Impacts, Cases of Sexual Harassment in Higher Education, and Protection Related to Sexual Harassment. These findings indicate that the LDA method is capable of identifying the main topics in sexual harassment comments and helping to organize unstructured social media data into information that is easier to understand. The contribution of this study is the proposed Latent Dirichlet Allocation (LDA)-based topic modeling approach with hyperparameter tuning to identify and organize unstructured sexual harassment comments on the social media platform X into coherent and interpretable topic clusters. The resulting topic mapping provides valuable insights into the issues that receive the greatest public attention and can serve as a foundation for understanding public concerns. Furthermore, these findings have the potential to support the development of victim support services, including telemedicine-based systems.

Keywords: Text Analysis; LDA; Social Media X; Sexual Harassment; Topic Modeling

1. PENDAHULUAN

Pelecehan seksual merupakan salah satu bentuk kekerasan berbasis *gender* yang berdampak pada kondisi psikologis, sosial, dan keamanan korban. Di Indonesia, isu kekerasan terhadap perempuan terus memperoleh perhatian publik seiring tingginya jumlah kasus yang tercatat. Berdasarkan Catatan Tahunan Komnas Perempuan 2025, terdapat 376.529 kasus kekerasan terhadap perempuan, dengan kekerasan seksual sebagai bentuk kekerasan yang paling banyak dilaporkan, yaitu sebesar 37.51% [1]. Data tersebut menunjukkan bahwa kekerasan seksual masih menjadi isu yang mendapat perhatian besar di ruang digital.

Media sosial memungkinkan terjadinya interaksi antarpengguna melalui komentar, dukungan, dan berbagi pengalaman pribadi. Dalam konteks pelecehan seksual, *platform* digital dapat menjadi ruang bagi korban untuk menyampaikan pengalaman, serta meningkatkan kesadaran publik terhadap isu kekerasan seksual. Salah satu *platform* yang banyak digunakan adalah X, karena menyediakan fitur komentar dan percakapan publik antarpengguna. Selain itu, pengguna tidak diwajibkan menggunakan identitas asli, sehingga *platform* ini dianggap lebih aman untuk berbagi pengalaman [2]. Komentar pada *platform* X juga dapat merepresentasikan respons publik terhadap isu pelecehan seksual.

Penelitian sebelumnya menggunakan analisis sentimen dan LIWC-22 untuk mengukur emosi masyarakat pada kasus pencabulan oleh Bechi, menunjukkan potensi data X sebagai sumber kajian terhadap isu pelecehan seksual [3]. Penelitian lainnya juga membahas peran media sosial dalam kasus kekerasan seksual anak untuk menjelaskan fungsi media sosial sebagai ruang advokasi publik sekaligus risiko reviktimisasi terhadap korban. Hal ini membuktikan bahwa media sosial dapat membantu meningkatkan perhatian publik terhadap kasus kekerasan seksual, tetapi juga dapat menimbulkan risiko baru bagi korban melalui komentar negatif, stigma, dan penyebaran identitas [4].



Besarnya keterlibatan masyarakat dalam membahas isu ini menghasilkan volume data percakapan yang terus meningkat di media sosial. Dengan demikian, banyaknya percakapan mengenai pelecehan seksual di media sosial X menunjukkan tingginya perhatian publik terhadap isu tersebut. Namun, volume data yang besar dan tidak terstruktur menyebabkan proses identifikasi topik pada kumpulan komentar secara manual menjadi sulit dilakukan. Sehingga penelitian ini menggunakan pendekatan pemodelan topik untuk mengidentifikasi topik utama dalam komentar pelecehan seksual di media sosial X.

Sejumlah penelitian sebelumnya menunjukkan bahwa media sosial X semakin banyak dimanfaatkan sebagai sumber data untuk memahami opini dan dinamika komentar publik. Diantaranya penelitian tentang deteksi pelecehan seksual dan predator obrolan media sosial, salah satunya penelitian mengenai deteksi pelecehan seksual pada percakapan digital menggunakan algoritma Naïve Bayes yang menunjukkan bahwa pendekatan berbasis teks dapat digunakan untuk mengidentifikasi konten yang mengarah pada pelecehan seksual [5]. Selain itu, penelitian tentang analisis jaringan pelecehan seksual digital lintas *platform* dengan menggunakan pendekatan *social network analysis* untuk melihat penyebaran isu pada media sosial X dan YouTube. Hasil penelitian menemukan bahwa isu pelecehan seksual digital dapat berkembang melalui interaksi lintas *platform* dan diperkuat oleh pola hubungan antar pengguna [6].

Media sosial X juga banyak digunakan dalam penelitian analisis sentimen untuk memahami opini dan respons masyarakat terhadap berbagai isu. Penelitian menggunakan metode Decision Tree C4.5 dan TF-IDF pada komentar Bitcoin menunjukkan bahwa data X dapat dimanfaatkan sebagai sumber informasi publik [7]. Penelitian lain menggunakan Support Vector Machine untuk menganalisis respons masyarakat melalui komentar mengenai layanan Samsat Digital Nasional [8]. Sementara itu, penelitian lain menggunakan Naïve Bayes dan SMOTE menunjukkan bahwa data X efektif digunakan untuk menganalisis opini publik pada isu politik [9]. Dengan demikian, media sosial X relevan digunakan sebagai sumber data untuk memahami bagaimana isu pelecehan seksual dibicarakan oleh masyarakat. Namun, data komentar yang diperoleh umumnya bersifat tidak terstruktur dan memiliki volume yang besar, sehingga diperlukan metode yang mampu mengidentifikasi pola pembahasan secara otomatis. Salah satu pendekatan yang banyak digunakan untuk tujuan tersebut adalah pemodelan topik (*topic modeling*), yang berfungsi untuk menemukan topik-topik utama dalam kumpulan data teks [10].

Dalam pemodelan topik, Latent Dirichlet Allocation (LDA) merupakan salah satu metode yang paling banyak digunakan untuk mengidentifikasi topik dominan pada data teks. Beberapa penelitian sebelumnya yang menggunakan LDA untuk menemukan topik dominan dalam komentar digital diantaranya adalah penelitian mengenai pemodelan topik pada ulasan hotel berbintang tinggi di Jakarta untuk menemukan topik utama terkait kualitas layanan, membuktikan bahwa model LDA dengan tujuh topik mendapatkan skor *coherence* 0,6065 mampu menggambarkan pola ulasan pelanggan, seperti keramahan staf, kualitas sarapan, kebersihan kamar, proses *check-in*, dan fasilitas hotel [11]. Penelitian lain pada laporan tugas akhir mahasiswa menunjukkan bahwa LDA mampu mengidentifikasi berbagai topik penelitian secara otomatis [12]. Selain itu, LDA juga digunakan untuk menganalisis tren topik penelitian tesis dan berhasil mengelompokkan data ke dalam beberapa topik utama [13].

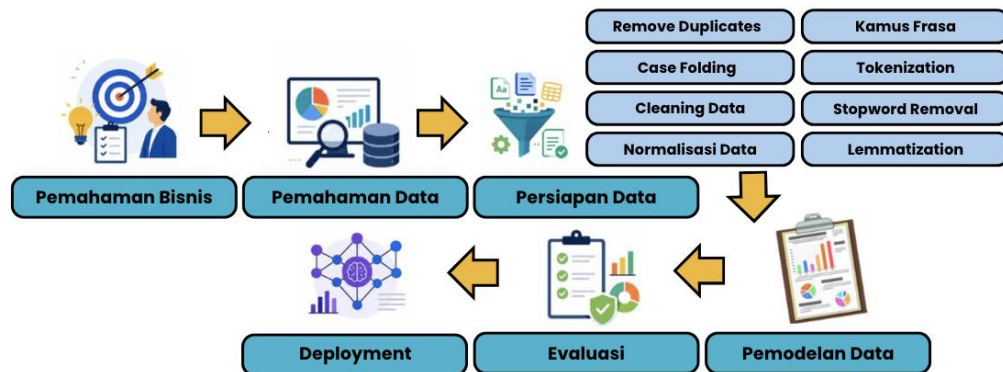
Meskipun penelitian sebelumnya menunjukkan bahwa LDA banyak digunakan untuk menemukan topik utama dari data teks tidak terstruktur, seperti ulasan hotel [11], laporan tugas akhir [12], dan tren topik penelitian [13], penelitian mengenai komentar publik terkait pelecehan seksual pada media sosial X masih terbatas. Belum banyak penelitian yang mengidentifikasi topik utama dari komentar publik terkait pelecehan seksual menggunakan pendekatan pemodelan topik. Komentar mengenai pelecehan seksual di media sosial juga memiliki kompleksitas linguistik tersendiri. Tindakannya sering tidak disebut langsung, melainkan melalui eufemisme seperti "tidak senonoh", kode seperti "KS", maupun penyamaran pelaku dengan sebutan "oknum", sehingga satu makna tersebar ke berbagai bentuk kata dan menyulitkan identifikasi topik bila pemodelan hanya bertumpu pada frekuensi kata.

Pendekatan berbasis transformer seperti BERTopic memang mampu menangani teks pendek melalui representasi kontekstual. Namun penelitian ini tetap memilih LDA karena hasilnya berupa distribusi probabilistik kata dan topik yang eksplisit sehingga lebih mudah ditafsirkan, di samping ketersediaan metrik evaluasi dan penelitian perbandingan berbahasa Indonesia yang memadai. Keterbatasan LDA pada teks pendek diatasi melalui normalisasi, penyusunan kamus frasa untuk menyatukan eufemisme dan kode menjadi satu token, serta lemmatization, sementara perbandingan dengan BERTopic terbuka sebagai arah penelitian selanjutnya. Berdasarkan pertimbangan tersebut, penelitian ini menggunakan LDA untuk mengidentifikasi topik utama pada komentar pelecehan seksual di media sosial X periode 2024–2026. Kebaruan penelitian terletak pada objek kajian berupa komentar publik terkait pelecehan seksual serta interpretasi topik berdasarkan kata kunci dominan, visualisasi, dan data representatif.

Tujuan utama penelitian ini adalah mengidentifikasi topik utama dalam komentar pelecehan seksual di media sosial X periode 2024–2026 melalui pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA). Dalam penelitian ini, data komentar dikelompokkan ke dalam beberapa topik yang direpresentasikan oleh kata-kata kunci dominan. Hasil pemodelan digunakan untuk menggambarkan topik yang paling banyak dibahas masyarakat terkait isu pelecehan seksual secara lebih terstruktur [14]. Kontribusi penelitian ini adalah menyajikan pendekatan pemodelan topik berbasis LDA dengan penyetelan *hyperparameter* untuk mengelompokkan komentar pelecehan seksual di media sosial X yang tidak terstruktur menjadi topik-topik utama yang lebih mudah dipahami. Hasil pemetaan topik ini diharapkan dapat menjadi dasar dalam memahami isu-isu yang paling banyak mendapat perhatian masyarakat serta berpotensi dimanfaatkan untuk mendukung pengembangan layanan dukungan korban, termasuk pada sistem telemedis

2. METODOLOGI PENELITIAN

Tahapan penelitian disajikan pada Gambar 1 menggunakan metode *Cross Industry Standard Process For Data Mining* (CRISP-DM) [15].



Gambar 1. Tahapan Penelitian

2.1 Objek Penelitian

Objek dalam penelitian ini adalah komentar pengguna pada media sosial X yang membahas isu pelecehan seksual. Data dikumpulkan melalui teknik *scraping* menggunakan kata kunci yang berkaitan dengan pelecehan seksual pada periode 2024–2026, dan dibatasi pada komentar berbahasa Indonesia. Setiap data terdiri atas teks komentar (*full_text*) beserta atribut pendukung seperti jumlah suka (*favorite_count*) dan bahasa (*lang*).

2.2 Pemahaman Bisnis

Pada tahap pemahaman bisnis, penelitian ini berawal dari tingginya volume jumlah komentar terkait pelecehan seksual pada media sosial X pada tahun 2024-2026. Komentar tersebut menggambarkan berbagai isu yang dibahas masyarakat, seperti pengalaman korban, dukungan sosial, kritik terhadap pelaku, serta respons terhadap penanganan kasus pelecehan seksual. Namun karena data komentar pada media sosial bersifat tidak terstruktur dan memiliki volume yang besar, sulit untuk mengetahui topik-topik yang sering dibahas oleh masyarakat mengenai topik pelecehan seksual. Oleh karena itu, penelitian ini bertujuan untuk menggunakan pemodelan topik *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik utama pada kumpulan komentar tersebut.

2.3 Pemahaman Data

Pada tahap ini melakukan pengumpulan data dengan metode *scraping* di media sosial X. Dataset yang terkumpul merupakan komentar terkait pelecehan seksual pada periode tahun 2024 hingga tahun 2026. Pencarian komentar menggunakan kata kunci “Pelecehan Seksual” dan “Kekerasan Seksual”. Total pengumpulan data 7.971 komentar. Berdasarkan hasil pengumpulan data, ditemukan bahwa komentar terbanyak terdapat pada tahun 2026 sebanyak 4,135 data, sedangkan tahun 2024 dan 2025 masing-masing berjumlah 1,971 dan 1,865 data. Dataset hasil pengumpulan tersebut memiliki beberapa atribut, namun penelitian ini hanya menggunakan atribut “*full_text*”, karena fokus penelitian adalah pemodelan topik berbasis teks. Atribut tersebut berisi kumpulan komentar pengguna tentang isu pelecehan seksual. Sedangkan atribut lain karena tidak terkait dengan topik penelitian, maka tidak digunakan. Contoh dataset pada kolom “*full_text*” ditampilkan pada Tabel 1.

Tabel 1. Sampel Dataset

No	full Text
1	Serendah-rendahnya tindakan kriminal adalah pelecehan seksual. universitas top kok pelecehan seksual, minimal itu ya ngacak-ngacak konstitusi laah.
2	Orang ga bakal tahu rasanya jadi korban pelecehan seksual kalau ga ngalamin langsung. Hanya karena laki terus bisa ngelawan atau teriak ke pelaku. Ini ceritaku sebagai korban pelecehan seksual. Kejadiannya saat ada dinas kantor. Aku sekamar bertiga. Both of them married men.
...	...
7971	Seorang perempuan yang mengalami lumpuh jadi korban pelecehan seksual sopir taksi online usai minta tolong dibantu jalan ke dalam rumah. https://t.co/giGXjyDTc

2.4 Persiapan Data

Pada tahap ini, dataset kemudian perlu dilakukan proses pembersihan data untuk memastikan kualitas data agar dapat digunakan secara optimal pada proses analisis [16]. Tahapan persiapan data dalam penelitian ini meliputi beberapa proses sebagai berikut:



- a. Penghapusan data duplikat adalah proses untuk membuang komentar dengan isi yang identik menggunakan “drop_duplicates”. Dari 7.971 komentar yang dikumpulkan, tersisa 6.288 komentar unik setelah proses ini. Penghapusan ini penting agar unggahan berulang seperti *retweet* atau penyalinan ulang komentar yang bersifat *spam* atau tidak memberi bobot berlebih pada topik tertentu.
- b. *Case Folding* adalah proses menyeragamkan bentuk huruf dengan mengubah seluruh teks menjadi huruf kecil sehingga format data menjadi konsisten. Contoh implementasi proses *case folding* di penelitian ini ditampilkan pada Tabel 2.

Tabel 2. Proses Case Folding

full text	text casefolding
Pantesan saja hukum kita banyak yang gak berpihak kepada korban pelecehan dan kekerasan seksual. Ya, mahasiswa hukumnya saja begini!?	pantesan saja hukum kita banyak yang gak berpihak kepada korban pelecehan dan kekerasan seksual ya mahasiswa hukumnya saja begini!?

- c. Pembersihan data adalah proses membersihkan teks dari elemen yang tidak dibutuhkan, seperti URL, mention, hashtag, angka, tanda baca, simbol, emoji, dan spasi berlebih dengan tujuan agar data teks menjadi lebih bersih sebelum dianalisis. Contoh implementasi proses *case folding* di penelitian ini ditampilkan pada Tabel 3.

Tabel 3. Proses Pembersihan data

text casefolding	text clean
kasus dugaan pelecehan viral di fh ui ! kasus dugaan kekerasan seksual verbal yang melibatkan mahasiswa fakultas hukum universitas indonesia (fh ui), universitas indonesia, tengah menjadi sorotan publik dan viral di media sosial sejak 11–12 april 2026. https://t.co/jdd81k46gm	kasus dugaan pelecehan viral di fh ui kasus dugaan kekerasan seksual verbal yang melibatkan mahasiswa fakultas hukum universitas indonesia fh ui universitas indonesia tengah menjadi sorotan publik dan viral di media sosial sejak april

- d. Normalisasi adalah proses mengubah kata tidak baku, singkatan, atau bahasa informal menjadi bentuk yang lebih baku. Contoh kamus normalisasi ditampilkan pada Tabel 4.

Tabel 4. Proses Normalisasi

Awal	normalisasi	Awal	normalisasi
cewe, cewek, cwek, perempuan	wanita	laki, cwok, cowo, cowok	pria
aku, gw, gua, gue, sy	saya	jg	juga
ga, nggak, nggk, gak, ngga, gk, tdk	tidak	yg	yang
tmn, temen, sahabat	teman	klo	kalau
ks	Kekerasan seksual	utk	untuk
bgimana, bgmn	bagaimana	support	dukung
Pd	pada	nitizen	masyarakat

- e. Kamus Frasa merupakan tahap menggabungkan frasa penting yang memiliki makna khusus agar tidak terpisah saat pemrosesan. Tahap ini bertujuan untuk membantu LDA menangkap makna istilah secara utuh. Contoh implementasi kamus frasa ditampilkan pada Tabel 5.

Tabel 5. Proses Kamus Frasa

frasa asli	Hasil penggabungan frasa
pelecehan seksual	pelecehan_seksual
kekerasan seksual	kekerasan_seksual
pelecehan verbal	pelecehan_verbal
kekerasan verbal	kekerasan_verbal
komentar seksual	komentar_seksual
tindakan asusila	tindakan_asusila
tindak pidana	tindak_pidana
cat calling	catcalling
media sosial	media_sosial
grup chat	grup_chat
kepala sekolah	kepala_sekolah
dokter kandungan	dokter_kandungan

- f. *Tokenization* adalah proses untuk memecah teks menjadi satuan kata atau token. Hasil dari tahap ini berupa daftar kata yang akan diproses pada tahap berikutnya. Contoh implementasi proses *tokenization* ditampilkan pada Tabel 6.

**Tabel 6.** Proses Tokenization

text normalized	tokens
pantesan saja hukum kita banyak yang tidak berpihak kepada korban pelecehan dan kekerasan seksual ya mahasiswa hukumnya saja begini	['pantesan', 'saja', 'hukum', 'kita', 'banyak', 'yang', 'tidak', 'berpihak', 'kepada', 'korban', 'pelecehan', 'dan', 'kekerasan_seksual', 'ya', 'mahasiswa', 'hukumnya', 'saja', 'begini']

- g. *Stopword Removal* bertujuan menghapus kata-kata umum yang tidak memiliki kontribusi signifikan terhadap pembentukan topik, seperti kata penghubung dan kata depan. Penelitian ini menggunakan daftar *Stopword* bahasa Indonesia dari pustaka NLTK serta tambahan *Stopword* yang disesuaikan dengan karakteristik dataset penelitian. Tahap ini membantu menghasilkan kata-kata yang lebih merepresentasikan topik utama komentar. Tabel 7 menampilkan contoh kamus *Stopword Removal* yang ditambahkan.

Tabel 7. Tambahan kamus *Stopword Removal*

Bahas	Remake	End	Tolol	Cuey	Iyahh	Selengkapnya	Tidur
Yaa	Beda	Trus	Kaya	Berusaha	Sebentar	Nemi	Dah
Sadar	Parah	Normal	Mending	Hai	Up	News	Kaum
Blow	Mostly	Group	Goblok	Has	Detail	Silenced	Yunus
Anjing	Lah	Kok	Tuh	Kalopoda	Ender	Sounds	Has
Anjg	Konten	Emg	Wkwk	Has	Kak	To	drpd
Iyahh	Intinyadeh	Btw	mcm	Pls	Bayangin	Apapun	Emg
Lu	Lo	Eth	Tw	Pilih	Segetol	Dijunjung	Prayers
Press	Tsb	Nanny	Thank	Story	Via	Jaringan	Peningkatan
Facts	Gigin	Bkn	Klu	Beritasatu	Dugaan	Mengundang	Ngambil
Cuey	Iyahh	Selengkapnya	Tidur	Breakfast	Gmn	Abi	Scene

- h. *Lemmatization* bertujuan mengubah kata ke bentuk dasarnya sesuai kamus. Tahap ini dilakukan agar kata-kata yang memiliki makna sama atau mirip dapat diseragamkan sehingga lebih mudah diproses oleh sistem [17]. Pada Tabel 8 menampilkan contoh implementasi proses *lemmatization*.

Tabel 8. Proses *Lemmatization*

kata awal	kamus lemma	kata awal	kamus lemma
korbannya	korban	ketakutan	takut
korban_kekerasan_seksual	korban	ditakuti	takut
korban_pelecehan_seksual	korban	kepolisian	polisi
pelakunya	pelaku	perempuan	wanita
pelaku_pelecehan_seksual	pelaku	cewe	wanita
pelaku_kekerasan_seksual	pelaku	cewek	wanita
perkosaan	pemeriksaan	laki_laki	pria
melapor	lapor	lelaki	pria
dilaporkan	lapor	cowok	pria
melaporkan	lapor	cowo	pria
penanganan	tangani	krl	transportasi
menangani	tangani	bis	transportasi
ditangani	tangani	melindungi	perlindungan
tertangani	tangani	dilindungi	perlindungan
traumatis	trauma	perlindungan	perlindungan

2.5 Pemodelan Data

Pada tahap pemodelan data, dataset hasil persiapan data digunakan sebagai input dalam pemodelan topik menggunakan LDA pada pustaka Gensim. Data dengan token kosong dihapus agar hanya data hasil persiapan data yang digunakan dalam pemodelan. Karena komentar pada media sosial X berupa teks pendek yang membatasi kemunculan bersama antar-kata, sebelum pemodelan diterapkan teknik penggabungan data untuk meningkatkan koherensi topik pada LDA untuk teks pendek. Selanjutnya, token pada setiap pseudo-data digunakan untuk membentuk *dictionary* dan *corpus*. *Dictionary* berisi daftar kata unik, sedangkan *corpus* merepresentasikan data dalam bentuk bag-of-words (BOW) berdasarkan frekuensi kemunculan kata. Setelah *dictionary* terbentuk, dilakukan penyaringan kata menggunakan parameter `no_below = 5` dan `no_above = 0,50`.

Parameter `no_below` menghapus kata yang muncul pada kurang dari 5 data, sedangkan `no_above` menghapus kata yang muncul pada lebih dari 50% data. Untuk mencegah hilangnya istilah inti yang berfrekuensi tinggi namun bermakna, sejumlah kata kunci penting dipertahankan melalui parameter `keep_tokens`. Hasil penyaringan menghasilkan 1.214 kata unik dan 953 pseudo-data dalam bentuk *corpus*. Setelah *dictionary* dan *corpus* terbentuk, model LDA dibangun



menggunakan `gensim.models.LdaModel`. Pemodelan dilakukan melalui hyperparameter tuning, yaitu pencarian kombinasi terbaik atas jumlah topik (K) serta parameter alpha (α) dan beta (β). Parameter alpha mengatur distribusi topik pada setiap data, sedangkan beta mengatur distribusi kata pada setiap topik.

Setiap kombinasi menghasilkan kumpulan topik yang direpresentasikan oleh kata-kata dominan beserta bobot probabilitasnya, kemudian dievaluasi menggunakan nilai coherence score (c_v). Kombinasi dengan nilai coherence tertinggi dipilih sebagai model akhir dengan tetap mempertimbangkan kemudahan interpretasi topik. Adapun hasil pemodelan LDA berupa kata kunci dominan, topik dominan pada setiap data, serta distribusi data berdasarkan topik disajikan pada bagian hasil dan pembahasan.

2.6 Evaluasi

Pada tahap evaluasi, nilai *coherence score* digunakan untuk mengukur keterkaitan antarkata dalam topik yang dihasilkan oleh model LDA. Semakin tinggi nilai *coherence*, semakin baik keterpaduan kata dalam topik tersebut, sehingga nilai tersebut digunakan sebagai dasar dalam menentukan jumlah topik terbaik [18]. Evaluasi dilakukan pada beberapa skenario jumlah topik menggunakan pustaka Gensim melalui `CoherenceModel` untuk menghitung nilai *coherence score*. Model dengan nilai *coherence score* tertinggi dipilih sebagai model terbaik.

Evaluasi dilakukan melalui *hyperparameter tuning*, yaitu pada beberapa kombinasi jumlah topik serta parameter alpha dan beta, menggunakan pustaka Gensim melalui `CoherenceModel` untuk menghitung nilai *coherence score* (c_v). Kombinasi dengan nilai *coherence score* tertinggi dipilih sebagai model terbaik. Selain evaluasi kuantitatif, penelitian ini juga menggunakan visualisasi `pyLDAvis` untuk membantu interpretasi hasil pemodelan. Pada bagian kiri visualisasi, setiap lingkaran merepresentasikan topik yang dihasilkan model, sedangkan jarak antar lingkaran menunjukkan tingkat kedekatan atau perbedaan antar-topik. Pada bagian kanan, ditampilkan kata-kata relevan beserta frekuensi kemunculannya pada topik yang dipilih. Visualisasi ini digunakan sebagai evaluasi kualitatif untuk memperkuat interpretasi topik yang dihasilkan [19]. Hasil evaluasi kemudian disajikan dalam bentuk tabel menggunakan pustaka `Pandas`.

2.7 Deployment

Pada tahap *deployment*, hasil pemodelan disajikan dalam bentuk ringkasan topik yang memuat kata kunci dominan, interpretasi topik, persentase kemunculan, dan penentuan nilai *coherence* yang didapatkan melalui *hyperparameter tuning*. Ringkasan ini digunakan untuk menggambarkan fokus pembahasan masyarakat terkait isu pelecehan seksual pada media sosial X khususnya periode 2024-2026.

3. HASIL DAN PEMBAHASAN

3.1 Penentuan Model Topik Terbaik

Penentuan model topik terbaik dilakukan melalui dua tahap. Tahap pertama adalah menentukan jumlah topik (K) yang paling optimal dengan membandingkan nilai coherence score (c_v) pada beberapa jumlah topik menggunakan parameter dasar (symmetric). Perbandingan tersebut disajikan pada Tabel 9.

Tabel 9. Perbandingan Nilai *Coherence Score*

num topics	coherence score
2	0.242
3	0.376
4	0.500
5	0.449
6	0.466

Berdasarkan Tabel 9, jumlah empat topik ($K=4$) menghasilkan nilai coherence tertinggi sebesar 0,500 dibandingkan jumlah topik lainnya, sehingga dipilih sebagai jumlah topik yang paling sesuai. Setelah jumlah topik ditentukan, dilakukan hyperparameter tuning terhadap parameter alpha (α) yang mengatur distribusi topik per data dan beta (β) yang mengatur distribusi kata per topik. Kombinasi terbaik diperoleh pada $\alpha=0,1$ dan $\beta=0,01$, yang meningkatkan nilai coherence menjadi 0,517. Konfigurasi inilah ($K=4$, $\alpha=0,1$, $\beta=0,01$) yang kemudian digunakan sebagai model final karena menghasilkan nilai evaluasi tertinggi sekaligus interpretasi topik yang paling jelas dan tidak tumpang tindih.

3.2 Visualisasi Model LDA

Pada tahap ini, ditampilkan visualisasi empat topik terbaik hasil pemodelan LDA dengan nilai *coherence score* tertinggi. Visualisasi dilakukan menggunakan Word Cloud dan `pyLDAvis`.

3.2.1 Visualisasi dengan Word Cloud

Word cloud digunakan untuk mengetahui kata-kata dominan pada setiap topik. Gambar 2 menampilkan visualisasi Word Cloud pada Topik 1. Kata-kata yang paling dominan pada topik ini antara lain "polda_metro", "rektor", "korban",

"pelecehan seksual", "lapor", "siswa", dan "kampus". Dominasi kata-kata tersebut menunjukkan bahwa Topik 1 berkaitan dengan penanganan kasus pelecehan seksual di lingkungan pendidikan.



Gambar 2. Word cloud pada Topik 1

Gambar 3 menampilkan visualisasi Word Cloud pada Topik 2. Kata-kata yang dominan pada topik ini antara lain "paksa", "sedih", "sakit", "nangis", "capek", "kuat", "peluk", dan "istirahat". Kemunculan kata-kata yang berkaitan dengan rasa emosional seseorang menunjukkan bahwa Topik 2 berkaitan dengan pengalaman dan dampak psikologis yang dialami korban.



Gambar 3. Word cloud pada Topik 2

Gambar 4 menampilkan visualisasi Word Cloud pada Topik 3. Kata-kata yang dominan pada topik ini antara lain "pria", "mahasiswa", "pelaku_pelecehan_seksual", "hukum", "media_sosial", "universitas_indonesia", "dosen", dan "grup_chat". Kombinasi kata-kata tersebut menunjukkan bahwa Topik 3 berkaitan dengan kasus pelecehan seksual di perguruan tinggi.



Gambar 4. Word cloud pada Topik 3

Gambar 5 menampilkan visualisasi Word Cloud pada Topik 4. Kata-kata yang dominan pada topik ini antara lain "lapor", "pelecehan seksual", "perempuan", "kantor", "trauma", "polisi", "transportasi", "kekerasan", dan "pria". Dominasi kata-kata tersebut menunjukkan bahwa Topik 4 berkaitan dengan perlindungan terkait pelecehan seksual.



Gambar 5. Word cloud pada Topik 4

3.2.2 Visualisasi pyLDAvis

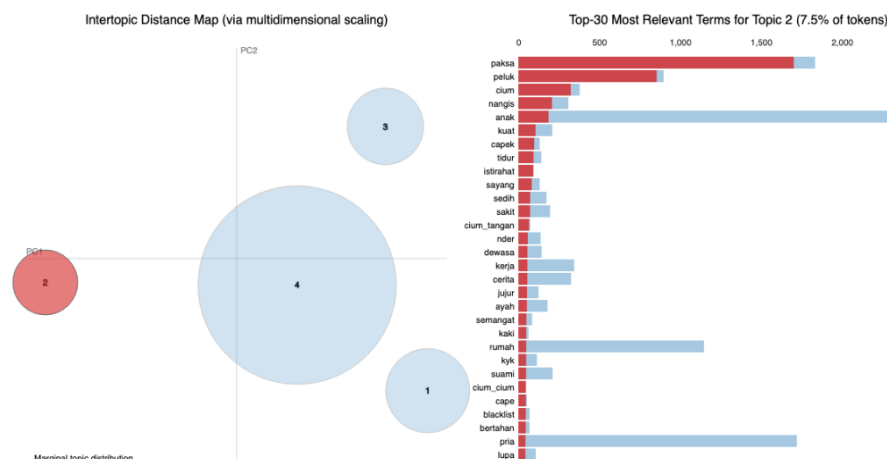
Pada tahap ini, hasil pemodelan LDA dengan 4 (empat) topik divisualisasikan menggunakan pyLDAvis. Visualisasi ini digunakan untuk menganalisis hubungan antar-topik serta mengidentifikasi kata-kata yang paling representatif pada setiap topik.

Gambar 6 menampilkan visualisasi pyLDAvis untuk Topik 1. Kata-kata yang paling relevan pada topik ini antara lain "polda_metro", "anak", "sekolah", "perlindungan", "keluarga", "rektor", "korban", "polisi", "lapur", "siswa", dan "guru". Kemunculan kata-kata yang berkaitan dengan ranah penegak hukum ("polisi", "polda_metro", "lapur") sekaligus lingkungan pendidikan ("sekolah", "siswa", "guru", "rektor") menunjukkan bahwa Topik 1 berkaitan dengan penanganan kasus pelecehan seksual di lingkungan pendidikan.



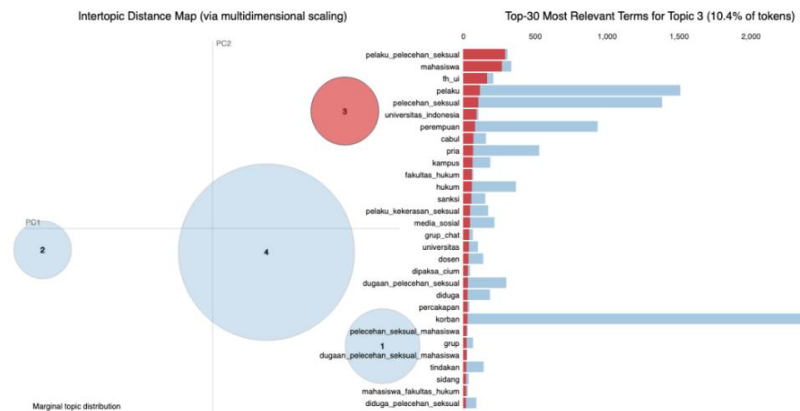
Gambar 6. Visualisasi pyLDAvis Topik 1

Gambar 7 menampilkan visualisasi pyLDAvis untuk Topik 2. Kata-kata yang paling relevan pada topik ini antara lain "paksa", "peluk", "cium", "nangis", "anak", "kuat", "capek", "tidur", "bertahan", dan "sedih". Kemunculan kata-kata yang sarat muatan emosional tersebut menunjukkan bahwa Topik 2 berkaitan dengan pengalaman dan dampak psikologis yang dialami korban.



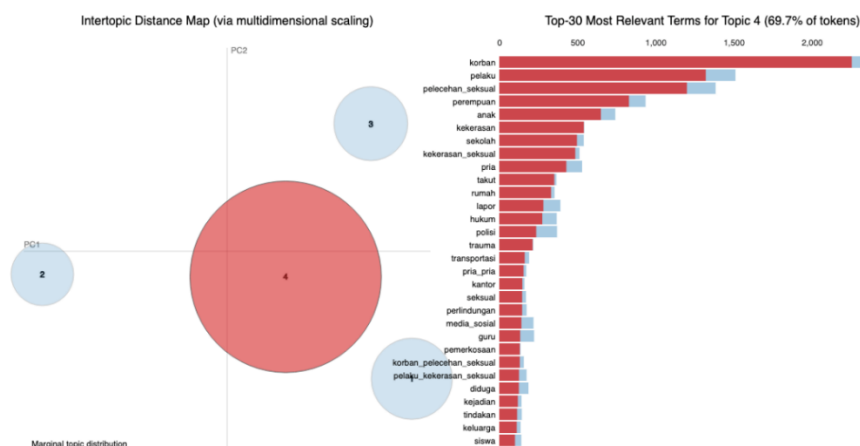
Gambar 7. Visualisasi pyLDAvis Topik 2

Gambar 8 menampilkan visualisasi pyLDAvis untuk Topik 3. Kata-kata yang paling relevan pada topik ini antara lain "pelaku_pelecehan_seksual", "universitas_indonesia", "fh_ui", "mahasiswa", "pria", "kampus", "fakultas_hukum", "grup_chat", "universitas", "grup", dan "dosen". Dominasi kata-kata yang berkaitan dengan institusi perguruan tinggi tersebut menunjukkan bahwa Topik 3 berkaitan dengan kasus pelecehan seksual di perguruan tinggi.



Gambar 8. Visualisasi pyLDAvis Topik 3

Gambar 9 menampilkan visualisasi pyLDAvis untuk Topik 4. Kata-kata yang paling relevan pada topik ini antara lain "korban", "pelaku", "kekerasan", "perlindungan", "kejadian", "tindakan", "keluarga", "lapor", "hukum", dan "polisi". Kombinasi kata-kata tersebut menunjukkan bahwa Topik 4 berkaitan dengan perlindungan terkait pelecehan seksual.



Gambar 9. Visualisasi pyLDAvis Topik 4

3.2 Ringkasan Topik, Kata Kunci, dan Interpretasi Topik

Tabel 10 terdiri dari beberapa kolom utama, yaitu Topik Interpretasi, Tokens %, Kata Kunci dan Nilai *Coherence*. Pada Kolom “Topik Interpretasi” berisi label atau nama topik yang diberikan berdasarkan kata-kata dominan dan makna pembahasan pada topik tersebut. Contohnya, Topik 1 diberi label Penanganan Kasus Pelecehan Seksual di Lingkungan Pendidikan, karena didominasi oleh kata seperti “dugaan_pelecehan_seksual”, “korban, polisi”, “universitas_pancasila”, “sekolah”, “siswa”. Sedangkan Kolom “Tokens %” menunjukkan proporsi kemunculan topik dalam keseluruhan data. Misalnya, nilai 50,3% pada Topik 4 menunjukkan bahwa topik tersebut memiliki proporsi terbesar dibandingkan topik lainnya. Sementara, kolom “Kata Kunci dan Nilai *Coherence*” memuat kata-kata paling relevan pada setiap topik beserta bobotnya. Contohnya, kata “korban” (7,53%) menunjukkan bahwa kata “korban” memiliki kontribusi paling besar dalam membentuk Topik 4.

Berdasarkan hasil pemodelan LDA, komentar mengenai pelecehan seksual di media sosial X terbagi ke dalam empat topik utama. Kumpulan kata dominan pada setiap topik kemudian diinterpretasikan menjadi “Penanganan Kasus Pelecehan di Lingkungan Pendidikan”. “Pengalaman dan Dampak Psikologis Korban”, “Kasus Pelecehan Seksual di Perguruan Tinggi”, serta “Perlindungan terkait Pelecehan Seksual”.

Tabel 10. Ringkasan Topik, Kata Kunci dan Interpretasi Topik

Topik Interpretasi	Tokens %	Kata Kunci dan Nilai <i>Coherence</i>
Penanganan Kasus Pelecehan Seksual di Lingkungan Pendidikan	20%	dugaan_pelecehan_seksual (4.96%), korban (2.72%), polisi (2.42%), universitas_pancasila (2.23%), berinisial (2.02%), lapor (1.70%), dosen (1.69%), guru (1.58%), jaya (1.45%), polda_metro (1.44%), pelaku (1.34%), pelecehan seksual (1.30%), periksa (1.02%), kampus (0.92%),



(Topik 1)		diduga_pelecehan_seksual (0.87%), rektor (0.82%), dokter (0.81%), mahasiswa (0.77%), sekolah (0.76%), siswa (0.72%), catcalling (0.67%), dugaan_pelecehan_seksual_universitas_pancasila (0.65%), hukum (0.58%), lapor_dugaan_pelecehan_seksual (0.57%), tersangka (0.53%), media sosial (0.51%), diduga (0.51%), mahkamah_goda (0.48%), pendidikan (0.47%), pelecehan_seksual_kantor (0.45%)
Pengalaman dan Dampak Psikologis Korban (Topic 2)	12.6%	paksa (16.33%), peluk (8.20%), cium (3.10%), nangis (1.99%), anak (1.79%), kuat (1.02%), capek (0.94%), tidur (0.90%), istirahat (0.89%), sayang (0.79%), sedih (0.70%), sakit (0.69%), cium_tangan (0.63%), nder (0.56%), dewasa (0.54%), kerja (0.54%), cerita (0.53%), jujur (0.51%), ayah (0.51%), semangat (0.48%), kaki (0.47%), rumah (0.46%), kyk (0.46%), suami (0.46%), cium_cium (0.44%), cape (0.44%), blacklist (0.44%), bertahan (0.44%), pria (0.42%), lupa (0.42%)
Kasus Pelecehan Seksual di Perguruan Tinggi (Topic 3)	17.0%	pelaku_pelecehan_seksual (6.56%), mahasiswa (6.05%), fh_ui (3.75%), pelaku (2.61%), pelecehan_seksual (2.38%), universitas_indonesia (2.11%), perempuan (1.86%), cabul (1.60%), pria (1.55%), kampus (1.47%), fakultas_hukum (1.38%), hukum (1.37%), sanksi (1.28%), pelaku_kekerasan_seksual (1.08%), media_sosial (1.07%), grup_chat (0.96%), universitas (0.91%), dosen (0.86%), dipaksa_cium (0.76%), dugaan_pelecehan_seksual (0.76%), diduga (0.71%), percakapan (0.71%), korban (0.70%), pelecehan_seksual_mahasiswa (0.57%), grup (0.56%), dugaan_pelecehan_seksual_mahasiswa (0.55%), tindakan (0.50%), sidang (0.49%), mahasiswa_fakultas_hukum (0.49%), diduga_pelecehan_seksual (0.46%)
Perlindungan Terkait Pelecehan Seksual (Topic 4)	50.3%	korban (7.53%), pelaku (4.41%), pelecehan_seksual (4.01%), perempuan (2.77%), anak (2.17%), kekerasan (1.81%), sekolah (1.66%), kekerasan_seksual (1.62%), pria (1.43%), takut (1.17%), rumah (1.11%), lapor (0.94%), hukum (0.92%), polisi (0.79%), trauma (0.71%), transportasi (0.54%), pria_pria (0.52%), kantor (0.49%), seksual (0.49%), perlindungan (0.49%), media_sosial (0.47%), guru (0.45%), pemerkosaan (0.44%), korban_pelecehan_seksual (0.44%), pelaku_kekerasan_seksual (0.42%), diduga (0.42%), kejadian (0.40%), tindakan (0.38%), keluarga (0.37%), siswa (0.33%)

Hasil pemodelan menunjukkan bahwa komentar mengenai pelecehan seksual di media sosial X tidak hanya membahas peristiwa kekerasan seksual, tetapi juga mencakup dampak psikologis terhadap korban, respons sosial dan institusional terkait penanganan kasus, serta perlindungan terkait terjadinya kekerasan seksual. Penelitian menunjukkan bahwa LDA dapat membantu mengelompokkan data komentar yang tidak terstruktur menjadi beberapa topik utama yang lebih mudah dianalisis. Selain itu, hasil pemodelan ini juga dapat digunakan untuk memahami isu yang paling banyak mendapat perhatian masyarakat, terutama terkait perlindungan korban, pelaporan kasus, dan pencegahan kekerasan seksual di berbagai lingkungan sosial.

3.3 Pembahasan

Berbeda dengan penelitian terdahulu yang lebih banyak memanfaatkan media sosial X untuk deteksi konten, klasifikasi, sentimen, atau analisis jejaring, misalnya dengan Naïve-Bayes untuk mendeteksi percakapan pelecehan seksual [5], social network analysis untuk melihat pola penyebaran lintas platform [6], serta SVM dan Naïve-Bayes untuk analisis respons publik pada konteks lain [8][9], penelitian ini berfokus menemukan tema-tema utama yang muncul dalam komentar publik di sosial media X mengenai pelecehan seksual melalui pendekatan Latent Dirichlet Allocation (LDA). Dengan demikian, penelitian ini tidak bertujuan mengklasifikasikan atau mendeteksi suatu jenis konten, melainkan menggali struktur topik yang merepresentasikan persepsi, perhatian, dan kebutuhan masyarakat terhadap isu pelecehan seksual. Sejalan dengan sejumlah penelitian mengenai pemodelan LDA sebelumnya yang berhasil mengidentifikasi topik dominan pada data teks tidak terstruktur [11][12][13], hasil penelitian ini menunjukkan bahwa komentar publik di X tidak hanya berfokus pada peristiwa kekerasan, tetapi juga pada dampak psikologis korban dan kebutuhan perlindungan. Temuan ini juga memberikan perspektif yang lebih komprehensif mengenai isu pelecehan seksual dibandingkan penelitian terdahulu yang berfokus pada aspek klasifikasi atau sentimen. Selain itu, penelitian ini menawarkan kontribusi pada aspek implementasi dalam pengembangan sistem telemedis, khususnya layanan konsultasi psikologis jarak jauh, sehingga keluhan dan kebutuhan korban dapat dikenali secara lebih terstruktur dan diarahkan pada penanganan yang tepat. Dengan demikian, model ini diharapkan tidak berhenti sebagai alat analisis, melainkan dapat membantu memperluas akses korban terhadap layanan kesehatan mental sekaligus mendukung upaya pencegahan kekerasan seksual.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa metode Latent Dirichlet Allocation (LDA) dapat digunakan untuk mengidentifikasi topik utama pada komentar pelecehan seksual di media sosial X pada periode 2024–2026. Setelah melalui tahap



pemahaman bisnis, pemahaman data, persiapan data, pemodelan topik, evaluasi, serta deployment, percakapan yang semula tersebar dalam bentuk komentar tidak terstruktur dapat dikelompokkan ke dalam beberapa topik utama. Hasil eksperimen menemukan bahwa model terbaik diperoleh pada empat topik dengan nilai coherence score sebesar 0,517. Penelitian ini memperlihatkan bahwa pembahasan masyarakat mengenai pelecehan seksual di media sosial X pada periode 2024–2026 tidak hanya berpusat pada peristiwa kekerasan itu sendiri, tetapi juga mencakup dampak terhadap korban, respons sosial dan institusional terkait penanganan kasus, serta lokasi atau ruang terjadinya kekerasan seksual. Temuan ini menjawab tujuan penelitian, yaitu mengidentifikasi topik utama pada komentar pelecehan seksual di media sosial X pada periode 2024–2026 menggunakan metode LDA. Kontribusi penelitian ini adalah menyajikan pendekatan pemodelan topik berbasis LDA dengan penyetelan *hyperparameter* untuk mengelompokkan komentar pelecehan seksual di media sosial X yang tidak terstruktur menjadi topik-topik utama yang lebih mudah dipahami. Hasil pemetaan topik ini diharapkan dapat menjadi dasar dalam memahami isu-isu yang paling banyak mendapat perhatian masyarakat serta berpotensi dimanfaatkan untuk mendukung pengembangan layanan dukungan korban, termasuk pada sistem telemedis. Dari sisi manfaat, hasil penelitian ini dapat memberikan gambaran awal mengenai hal-hal yang menjadi perhatian publik dalam isu pelecehan seksual. Informasi tersebut dapat digunakan sebagai bahan pertimbangan bagi masyarakat, lembaga pendidikan, pemerintah, maupun pihak terkait dalam memperkuat edukasi, perlindungan korban, mekanisme pelaporan, dan penanganan kasus. Penelitian selanjutnya disarankan untuk menempuh beberapa langkah teknis, seperti meningkatkan koherensi melalui penggabungan data (*document pooling*) berdasarkan kemiripan atau tagar, memperluas dan memperdalam tahap *pre-processing*, khususnya normalisasi kata tidak baku serta penyusunan frasa agar istilah bermakna tidak terpecah, melakukan penyetelan *hyperparameter* terhadap jumlah topik, α , dan β secara sistematis, serta memperluas sumber dan jumlah data agar hasil pemodelan lebih optimal. Selain itu, metode LDA dapat dibandingkan dengan NMF, BERTopic, atau Top2Vec untuk melihat perbedaan kualitas topik yang dihasilkan.

REFERENCES

- [1] Komnas Perempuan, “Siaran Pers Komnas Perempuan Peluncuran Catatan Tahunan Kekerasan terhadap Perempuan 2025,” <https://komnasperempuan.go.id/>, 2026.
- [2] R. Dwianatha Putri and R. Hermawati, “Pemanfaatan Media Sosial Twitter Sebagai Ruang Bercerita Bagi Korban Pelecehan Seksual,” *Jurnal Ilmu Sosial dan Ilmu Politik Malikussaleh (JSPM)*, vol. 5, no. 2, pp. 18–35, 2024, doi: 10.29103/jspm.v5i2.14863.
- [3] H. D. Faiqal, A. D. Gita, B. A. R. Zaki, and T. B. Haidar, “Psikologi siber: Reaksi warganet Twitter terhadap kasus pencabulan oleh Mas Bechi sebagai cerminan nilai dan sikap,” *Jurnal Psikologi Sosial*, vol. 22, no. 1, pp. 41–53, 2024, doi: 10.7454/jps.2024.06.
- [4] D. A. Puannandini, D. Anggraeni, D. Octo Firmansyah Gulo, A. Ahmad Nur, and J. K. Karabi, “Peran Ganda Media Sosial Dalam Kasus Kekerasan Seksual Anak,” *Adagium: Jurnal Ilmiah Hukum*, vol. 3, no. 2, pp. 387–398, 2025, doi: 10.70308/adagium.v3i2.230.
- [5] M. A. F. Riyadi, E. K. Andana, and M. A. Haq, “Deteksi Pelecehan Seksual dan Predator Obrolan Media Sosial Menggunakan Naive Bayes,” in *Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika (SNESTIK)*, 2024, pp. 47–51. doi: 10.31284/p.snestik.2024.5882.
- [6] L. Basit, P. Santoso, and F. Rizky, “Multi-platform analysis of sexual harassment networks: gender dynamics and digital amplification,” *Soc. Neww. Anal. Min.*, vol. 16, no. 1, pp. 16–18, 2026, doi: 10.1007/s13278-025-01563-3.
- [7] H. P. Indrizal, F. Syafria, E. Haerani, Y. Vitriani, and Y. Yusra, “Klasifikasi Sentimen Bitcoin Terhadap Komentar Di Aplikasi X Menggunakan Metode Decision Tree C4.5,” *Bulletin of Computer Science Research*, vol. 6, no. 1, pp. 469–478, 2025, doi: 10.47065/bulletincsr.v6i1.932.
- [8] A. Sasi Kirana, Rusdah, R. Roeswidiah, and A. Pudoli, “Analisis Sentimen Pada Media Sosial Terhadap Layanan SAMSAT Digital Nasional Dengan Support Vector Machine,” *Idealis: Indonesia Journal Information System*, vol. 8, no. 1, pp. 53–63, 2025, doi: 10.36080/idealis.v8i1.3276.
- [9] F. Nur Salsabilla and A. Witanti, “Analisis Sentimen Akhir Masa Jabatan Presiden Jokowi Pada Media Sosial X Menggunakan Naive Bayes,” *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 8, no. 1, pp. 106–115, 2025, doi: 10.36080/skanika.v8i1.3331.
- [10] M. Hankar, M. Kasri, and A. Beni-Hssane, “A comprehensive overview of topic modeling: Techniques, applications and challenges,” *Neurocomputing*, vol. 628, pp. 129638, 2025, doi: 10.1016/J.NEUCOM.2025.129638.
- [11] T. Mandiri and S. Juanita, “Discovering Service Quality from High-Rated Hospitality Reviews in Jakarta Using LDA,” in *2026 International Seminar on Intelligent Business and Edge-Computing Research (ISIBER)*, Jakarta, Indonesia: IEEE, 2026, pp. 868–873. doi: 10.1109/ISIBER68248.2026.11469849.
- [12] Z. Rosadi and A. Solichin, “Topic Modeling Tugas Akhir Mahasiswa Menggunakan Metode Latent Dirichlet Allocation Dengan Gibbs Sampling,” *Jurnal TICOM: Technology of Information and Communication*, vol. 13, no. 1, pp. 38–44, 2024, doi: 10.70309/ticom.v13i1.140.
- [13] L. Bayuaji and A. Wahyudi, “Analisis Trend Topik Penelitian Tesis Pada Program Studi Magister Ilmu Komputer Universitas Budi Luhur Menggunakan Metode Latent Dirichlet Allocation (LDA),” *Faktor Exacta*, vol. 17, no. 1, pp. 77–84, 2024, doi: 10.30998/faktorexacta.v17i1.21190.
- [14] O. S. Nufi and Khalid, “Topic Modelling Skripsi Manajemen Dakwah PTKIN menggunakan Latent Dirichlet Allocation (LDA),” *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 9, no. 2, pp. 539–549, 2025, doi: 10.33395/remik.v9i2.14671.
- [15] A. A. Hasna and G. Hendratomo, “Cancel culture pelaku pelecehan seksual di media sosial,” *Dimensia: Jurnal Kajian Sosiologi*, vol. 13, no. 1, pp. 47–58, 2024, doi: 10.21831/dimensia.v13i1.60990.
- [16] D. B. Saputra, V. Atina, and F. E. Nastiti, “Penerapan Model CRISP-DM Pada Prediksi Nasabah Kredit Menggunakan Algoritma Random Forest,” *Idealis: Indonesia Journal Information System*, vol. 7, no. 2, pp. 240–247, 2024, doi: 10.36080/idealis.v7i2.3244.



- [17] D. S. Devianto, D. Rizqi Daifullah, M. Abdurrahman Assidiqi, M. Faiz Nabil Hasmi, and P. Keuangan Negara STAN, “Analisis Faktor-Faktor Yang Mempengaruhi Kinerja Perusahaan IDXBUMN20 Dengan Pendekatan Data Mining,” *Integrative Perspectives of Social and Science Journal*, vol. 2, no. 3, pp. 5998–6002, 2025.
- [18] A. Muhammad and R. A. Rasheed, “Machine Learning Models For Hausa-Based Language (Words) Lemmatization,” *FUDMA Journal of Sciences (FJS)*, vol. 9, no. 12, pp. 352–357, 2025, doi: 10.33003/fjs-2025-0912-4396.
- [19] E. Puspita, D. F. Shiddieq, and F. F. Roji, “Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc),” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 481–489, 2024, doi: 10.57152/malcom.v4i2.1204.
- [20] Y. Prastyo, W. Yustanti, and Y. Yamasari, “Uncovering Hidden Themes in Audit Findings Through LDA-Based Topic Modeling,” *Journal Information Engineering and Educational Technology*, vol. 9, no. 1, pp. 28–35, 2025, doi: 10.26740/jieet.v9n1.p28-35.