Volume 6 No. 1 (January - June 2026), pp. 42-64 | ISSN 2962-2360 (online) | ISSN 2962-3669 (print)
Arcitech: Journal of Computer Science and Artificial Intelligence
https://journal.iaincurup.ac.id/index.php/arcitech/article/view/16764 | DOI : https://dx.doi.org/10.29240/arcitech.v6i1.16764

Analisis Komparatif Pemodelan Topik Promosi Judi Online pada Komentar YouTube Menggunakan Latent Dirichlet Allocation dan BERTopic

Nur Aisyah Wahyuni[1], Hafiz Irsyad[2]
[1,2] Program Studi Informatika, Universitas Multi Data Palembang, Palembang, Indonesia
Email : nuraisyahwah@mhs.mdp.ac.id, hafizirsyad@mdp.ac.id

Article Information

Article history

Received 29 April 2026
Revised 25 May 2026
Accepted 4 June 2026
Available 05 June 2026

Keywords

Topic Modeling
Latent Dirichlet Allocation (LDA)
BERTopic
Online Gambling
Youtube

Corresponding Author:

Nur Aisyah Wahyuni
Program Studi Informatika, Universitas Multi Data Palembang
Email:
nuraisyahwah@mhs.mdp.ac.id

Abstract

This study aims to analyze topics in YouTube comments related to online gambling using Latent Dirichlet Allocation (LDA) and BERTopic, as well as to compare the performance of both methods. The dataset consists of 6,350 YouTube comments obtained from Kaggle. The analysis process includes preprocessing, topic modeling, and evaluation using topic coherence and topic diversity metrics. The results show that LDA achieves a topic coherence score of 0.511 and a topic diversity score of 1.0, while BERTopic achieves a topic coherence score of 0.667 and a topic diversity score of 0.449. These findings indicate that BERTopic produces more semantically coherent topics compared to LDA, although it has a higher level of overlap between topics. Furthermore, the interpretation results reveal that several identified topics are related to online gambling promotion, while others are influenced by noise in the comment data. Therefore, BERTopic is considered more effective for analyzing short and unstructured text data.

Keywords : *Topic Modeling, LDA, BERTopic, Online Gambling, YouTube*

Abstrak

Penelitian ini bertujuan untuk menganalisis topik pada komentar YouTube terkait judi online menggunakan metode *Latent Dirichlet Allocation* (LDA) dan BERTopic serta membandingkan performa kedua metode tersebut. Data yang digunakan berupa 6.350 komentar YouTube yang diperoleh dari Kaggle. Proses analisis meliputi tahap pre-processing, pemodelan topik, serta evaluasi menggunakan metrik *topic coherence* dan *topic diversity*. Hasil penelitian menunjukkan bahwa metode LDA menghasilkan nilai *topic coherence* sebesar 0.511 dan *topic diversity* sebesar 1.0, sedangkan BERTopic menghasilkan nilai *topic coherence* sebesar 0.667 dan *topic diversity* sebesar 0.449. Hasil tersebut menunjukkan bahwa BERTopic mampu menghasilkan topik yang lebih koheren secara semantik dibandingkan LDA, meskipun memiliki tingkat overlap antar topik yang lebih tinggi. Selain itu, hasil interpretasi menunjukkan bahwa beberapa topik yang terbentuk berkaitan dengan promosi judi online, sementara topik lainnya dipengaruhi oleh noise pada data komentar. Dengan demikian, BERTopic dinilai lebih efektif dalam menganalisis data teks pendek dan tidak terstruktur.

Kata Kunci : *Topic Modeling, LDA, BERTopic, Judi Online, YouTube*

1. Pendahuluan

Media sosial telah menjadi bagian yang tidak terpisahkan dari kehidupan masyarakat modern. Platform seperti YouTube, Instagram, X, dan TikTok tidak hanya dimanfaatkan sebagai sarana komunikasi dan hiburan, tetapi juga sebagai media penyebaran informasi secara luas dan cepat (David et al., 2017). Perkembangan media sosial memberikan dampak positif dalam penyebaran informasi dan aktivitas pemasaran digital. Namun, di sisi lain, perkembangan ini juga membuka peluang terhadap munculnya konten ilegal, seperti promosi judi online yang dapat memberikan dampak negatif bagi masyarakat (Irawan, 2024).

Judi online merupakan salah satu bentuk kejahatan digital yang mengalami peningkatan signifikan di Indonesia. Aktivitas ini memungkinkan pengguna melakukan taruhan secara daring dengan menggunakan uang asli melalui berbagai jenis permainan seperti taruhan olahraga, kasino virtual, dan permainan kartu (Vigar et al., 2019). Berdasarkan laporan Pusat Pelaporan dan Analisis Transaksi Keuangan (PPATK), nilai transaksi judi online di Indonesia mencapai Rp. 327 triliun, pada tahun 2023 dan meningkat menjadi Rp. 359.81 triliun, hingga tahun 2024 sebelum mengalami penurunan sebesar 20% tahun 2025 menjadi Rp. 286.84 triliun (Grehenson, 2024; PPATK, 2026). Fenomena ini menjadi ancaman serius karena melibatkan berbagai kalangan, termasuk remaja dan mahasiswa.

Sebagai salah satu platform media sosial dengan jumlah pengguna terbesar, YouTube sering dimanfaatkan sebagai sarana penyebaran konten promosi judi online. Promosi tersebut umumnya dilakukan secara terselubung melalui penggunaan *caption*, *hashtag*, akun anonim, maupun komentar spam sehingga sulit dideteksi secara manual. Selain itu, platform ini juga dimanfaatkan sebagai media pemasaran tidak langsung yang menargetkan pengguna muda dengan minimnya regulasi yang membatasi promosi tersebut (Gunadi & Sugiantari, 2024). Kondisi ini menimbulkan tantangan dalam upaya pengawasan dan penanggulangan konten ilegal di media sosial.

Dalam konteks analisis data, pendekatan berbasis text mining dan machine learning dapat digunakan untuk mengekstraksi informasi serta mengidentifikasi pola dan struktur tersembunyi dalam data teks (Jelita, 2024). Salah satu metode yang umum digunakan adalah topic modeling, yaitu teknik untuk mengelompokkan dokumen berdasarkan topik tertentu. Metode *Latent Dirichlet Allocation* (LDA) merupakan metode klasik yang digunakan dalam pemodelan topik dan terbukti mampu mengidentifikasi topik dominan dalam data teks (Blei et al., 2003; Syaifuddin et al., 2021).

Berdasarkan penelitian Syaifuddin et al. (2021), penerapan LDA dalam mengekstraksi topik pada percakapan media sosial menghasilkan nilai *precision* sebesar 0.9294, *recall* sebesar 0.7900, dan *f-measure* sebesar 0.8541, yang menunjukkan kinerja yang baik dalam analisis topik. Selain itu, Kannitha et al. (2022) melaporkan bahwa LDA menghasilkan nilai *topic coherence* sebesar 0.49, yang menunjukkan kemampuan metode ini dalam mengidentifikasi struktur topik secara efektif. Penelitian dari Handayani (2026)

menunjukkan bahwa penerapan *Latent Dirichlet Allocation* (LDA) dalam pemodelan topik berita online Indonesia menghasilkan nilai *topic coherence* sebesar 0.5600, *topic diversity* sebesar 0.8400, serta *silhouette score* sebesar 0.0229. Selain itu, waktu pelatihan model LDA mencapai 86.11 detik, yang menunjukkan bahwa metode ini mampu mengidentifikasi distribusi topik dalam dokumen, meskipun masih memiliki keterbatasan dalam efisiensi komputasi dan kualitas koherensi topik.

Penelitian lain oleh Nugraha & Utami (2024) juga menunjukkan bahwa LDA menghasilkan nilai *topic coherence* sebesar 0.30, yang mengindikasikan bahwa metode ini masih memiliki keterbatasan dalam menangkap hubungan semantik secara mendalam. Namun, LDA memiliki keterbatasan dalam menangani teks pendek dan tidak terstruktur karena kurang mampu menangkap konteks semantik secara mendalam (Nugraha & Utami, 2024).

Untuk mengatasi keterbatasan tersebut, dikembangkan metode yang lebih adaptif terhadap konteks semantik, yaitu BERTopic. BERTopic merupakan metode berbasis *transformer* yang memanfaatkan representasi embedding teks untuk menghasilkan topik yang lebih kontekstual (Grootendorst, 2022). Penelitian sebelumnya menunjukkan bahwa BERTopic memiliki nilai *topic coherence* yang lebih tinggi dibandingkan metode tradisional. Penelitian dari Nugraha & Utami (2024), menunjukkan bahwa BERTopic menghasilkan nilai *topic coherence* sebesar 0.53, lebih tinggi dibandingkan LDA sebesar 0.30.

Selain itu, penelitian Nursyahrina et al. (2024) juga menunjukkan bahwa BERTopic mencapai nilai *topic coherence* sebesar 0.63, yang menunjukkan kualitas topik yang lebih baik secara semantik. Pada penelitian Nanayakkara & Thennakoon (2024) menunjukkan bahwa BERTopic memiliki kinerja yang lebih baik dibandingkan LDA dan NMF dalam analisis komentar YouTube, dengan nilai *coherence* C_V sebesar 0.5056 dan nilai U_MASS sebesar −14.3149, sehingga menghasilkan topik yang lebih kohesif dan relevan secara semantik.

Dalam bidang NLP, topic modeling merupakan salah satu pendekatan yang banyak digunakan untuk mengidentifikasi struktur tema tersembunyi pada kumpulan dokumen teks. *Latent Dirichlet Allocation* (LDA) merupakan metode probabilistik klasik yang mampu memodelkan distribusi topik berdasarkan kemunculan kata dalam dokumen. Metode ini telah banyak digunakan dalam analisis media sosial dan terbukti efektif dalam menemukan pola topik dominan. Namun demikian, LDA memiliki keterbatasan dalam menangani data teks pendek dan tidak terstruktur karena pendekatan berbasis bag-of-words kurang mampu menangkap hubungan semantik dan konteks antar kata secara mendalam.

Untuk mengatasi keterbatasan tersebut, berkembang pendekatan topic modeling modern berbasis transformer seperti BERTopic. BERTopic memanfaatkan representasi embedding dari *model transformer* untuk menghasilkan pemodelan topik yang

lebih kontekstual dan semantik. Dengan memanfaatkan *Sentence*-BERT, UMAP, dan HDBSCAN, BERTopic mampu mengelompokkan dokumen berdasarkan kemiripan makna, bukan hanya frekuensi kata. Beberapa penelitian sebelumnya menunjukkan bahwa BERTopic menghasilkan *topic coherence* yang lebih baik dibandingkan metode probabilistik tradisional seperti LDA, khususnya pada data media sosial yang bersifat noisy dan short text

Meskipun berbagai penelitian telah membahas penerapan LDA dan BERTopic dalam *topic modeling,* sebagian besar penelitian masih berfokus pada data umum seperti berita, X, atau dokumen formal. Penelitian yang secara khusus mengevaluasi kemampuan *semantic topic modeling* dalam menganalisis komentar YouTube terkait promosi judi online masih sangat terbatas. Selain itu, karakteristik komentar YouTube yang mengandung spam, bahasa informal, dan *noise* semantik masih menjadi tantangan bagi metode topic modeling tradisional. Oleh karena itu, diperlukan evaluasi komparatif yang lebih mendalam untuk memahami efektivitas metode probabilistik dan *transformer-based topic modeling* dalam menangkap representasi topik pada data komentar media sosial yang kompleks.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk melakukan analisis dan evaluasi komparatif antara *Latent Dirichlet Allocation* (LDA) dan BERTopic dalam memodelkan topik pada komentar YouTube terkait promosi judi online. Evaluasi dilakukan menggunakan metrik *topic coherence* dan *topic diversity* untuk mengukur kualitas semantik dan keberagaman topik yang dihasilkan. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan semantic *topic modeling* pada data *short text* media sosial serta mendukung pengembangan sistem monitoring dan deteksi konten ilegal berbasis *Natural Language Processing*.

2. Kajian Terdahulu

2.1 Judi Online

Judi online merupakan aktivitas perjudian yang dilakukan melalui jaringan internet dengan menggunakan uang asli dan dapat diakses melalui berbagai perangkat digital seperti komputer dan ponsel pintar (Vigar et al., 2019). Perkembangan teknologi informasi dan meningkatnya penggunaan internet menyebabkan praktik judi online semakin meluas di masyarakat. Di Indonesia, judi online termasuk tindakan pidana yang dilarang berdasarkan Undang-Undang Informasi dan Transaksi Elektronik (ITE) Pasal 27 Ayat 2 serta Pasal 303 KUHP, dengan ancaman hukuman penjara dan denda (Husain, 2024). Upaya penanggulangan dilakukan oleh Kepolisian Republik Indonesia dan Kementerian Komunikasi dan Informatika melalui pemblokiran ribuan situs judi online setiap tahunnya (Sri Gustina et al., 2025).

Dalam bidang teknologi informasi, data terkait judi online dapat dianalisis untuk mendeteksi konten ilegal dan mengidentifikasi pola teks yang berkaitan dengan aktivitas

tersebut. Metode *machine learning* seperti *Support Vector Machine* (SVM) banyak digunakan untuk melakukan klasifikasi dan pengenalan pola dalam data teks (Samuel & Kristiadi, 2024). Selain itu, analisis jaringan dimanfaatkan untuk memetakan hubungan antar akun dalam jaringan perjudian daring. Meskipun penelitian sebelumnya mampu melakukan deteksi otomatis, masih terdapat keterbatasan pada jenis data dan metode yang digunakan. Oleh sebab itu, diperlukan pengembangan pendekatan yang lebih komprehensif dengan mengombinasikan berbagai metode analisis seperti text mining, machine learning, dan analisis jejaring sosial untuk meningkatkan akurasi deteksi aktivitas judi online.

2.2 YouTube

YouTube merupakan salah satu platform media sosial yang dapat membagikan video secara online, dengan tujuan sebagai sarana berbagi, mencari, dan menonton video yang dapat diakses dari berbagai belahan dunia melalui aplikasi maupun situs web (David et al., 2017). YouTube tidak hanya menjadi media hiburan, tetapi juga sebagai sarana penyebaran opini yang dapat mempengaruhi persepsi publik. Oleh karena itu, YouTube dimanfaatkan sebagai sumber data dalam menganalisis dinamika sosial, termasuk isu perjudian online.

Penelitian sebelumnya menunjukkan bahwa YouTube memiliki peran dalam penyebaran konten terkait judi online. Chamil et al. (2024) menyatakan bahwa YouTube digunakan sebagai media untuk menyampaikan konten yang membahas dampak psikologis dan sosial dari perjudian online, salah satunya melalui podcast Kemencast. Sementara itu, penelitian dari Gunadi & Sugiantari (2024) menemukan bahwa YouTube sering digunakan sebagai media pemasaran tidak langsung yang menargetkan kaum muda dengan minimnya regulasi. Selain itu, Media sosial berfungsi sebagai sarana yang menormalisasi praktik perjudian online melalui teknik soft selling, penggunaan bahasa tidak langsung, serta visual yang menarik (Indra & Srihadiati, 2025).

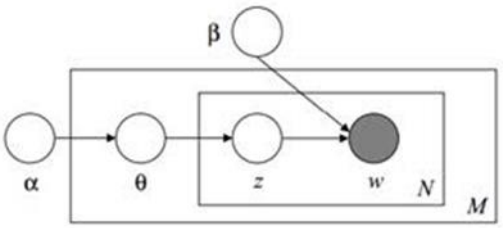
2.3 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) merupakan metode pemodelan topik berbasis probabilistik yang digunakan untuk mengidentifikasi topik tersembunyi dalam kumpulan dokumen teks. Metode ini mengasumsikan bahwa setiap dokumen terdiri dari beberapa topik, dan setiap topik direpresentasikan oleh distribusi kata tertentu (Blei et al., 2003).

Secara umum, probabilitas kemunculan kata dalam dokumen dapat direpresentasikan sebagai kombinasi distribusi topik dan distribusi kata, yang dirumuskan sebagai berikut:

$$p(w,d) = p(d) \sum_z p(z \mid d) p(w \mid z) \quad \text{.....(1)}$$

Persamaan tersebut menunjukkan bahwa suatu kata dalam dokumen dihasilkan dari distribusi topik tertentu, di mana setiap dokumen memiliki proporsi topik yang berbeda dan setiap topik memiliki distribusi kata yang berbeda. Dalam penerapannya, *model Latent Dirichlet Allocation* (LDA) dipengaruhi oleh sejumlah parameter utama, yaitu alpha (α) dan beta (β). Parameter alpha berfungsi mengatur distribusi topik dalam dokumen, sedangkan beta mengatur distribusi kata pada setiap topik. Selain itu, jumlah iterasi selama proses pelatihan juga mempengaruhi kestabilan dan kualitas topik yang dihasilkan. Model ini menggunakan distribusi *Dirichlet* sebagai prior untuk mengatur distribusi topik pada dokumen dan distribusi kata pada topik.



Gambar 1. Model Representasi LDA
Sumber: (Faizah dan Lin, 2023)

Gambar 1. menunjukkan hubungan antara dokumen, topik, dan kata dalam model LDA. Setiap dokumen memiliki distribusi topik (θ), sedangkan setiap topik memiliki distribusi kata (φ). Variabel z merepresentasikan topik tersembunyi yang menghasilkan kata dalam dokumen. Parameter α dan β merupakan parameter Dirichlet yang mengatur distribusi topik dan distribusi kata (Blei et al., 2003; Faizah & Lin, 2023).

Dalam penelitian ini, LDA digunakan sebagai metode dasar dalam pemodelan topik yang kemudian dibandingkan dengan metode BERTopic untuk menganalisis kualitas topik berdasarkan nilai *topic coherence* dan *topic diversity* pada data komentar YouTube terkait judi online.

2.4 BERTopic

BERTopic merupakan metode pemodelan topik modern yang memanfaatkan representasi semantik dari model transformer seperti BERT untuk menghasilkan embedding teks. Metode ini mampu menangkap makna kontekstual dari kata dan kalimat sehingga efektif digunakan untuk menganalisis teks pendek dan tidak terstruktur seperti komentar media sosial (Grootendorst, 2022).

Proses BERTopic terdiri dari beberapa tahapan utama, yaitu pembentukan *embedding* menggunakan Sentence-BERT, reduksi dimensi menggunakan UMAP, serta pengelompokan dokumen menggunakan algoritma *clustering* seperti HDBSCAN. Setelah itu, dilakukan ekstraksi kata kunci pada setiap topik menggunakan metode *class-based* TF-IDF (c-TF-IDF) untuk memperoleh representasi topik yang lebih jelas (Nursyahrina et al., 2024).

Dalam penerapannya, hasil BERTopic dipengaruhi oleh pengaturan parameter pada setiap komponen. Parameter seperti jumlah tetangga *(n_neighbors)* pada UMAP dan ukuran cluster minimum (*min_cluster_size*) pada HDBSCAN berperan dalam menentukan jumlah topik serta kualitas representasi topik (Grootendorst, 2022).

Secara matematis, bobot kata dalam suatu topik pada BERTopic dapat dihitung menggunakan pendekatan c-TF-IDF sebagai berikut:

$$W_{x,c} = \frac{tf_{x,c}}{||tf_c||} \times log(1 + \frac{A}{f_x}) \text{ ….. (2)}$$

Di mana $tf_{x,c}$ merupakan frekuensi kemunculan kata dalam klaster, $||tf_c||$ adalah total frekuensi kata dalam klaster, f_x adalah frekuensi kata dalam seluruh dokumen, dan A adalah rata-rata jumlah kata per klaster.

Dalam penelitian ini, BERTopic digunakan sebagai metode pembanding terhadap LDA untuk menganalisis kualitas topik berdasarkan nilai *topic coherence* dan *topic diversity* pada data komentar YouTube terkait judi online.

2.5 Analisis Penelitian Terdahulu, Research Gap, dan Kontribusi Penelitian

Penelitian mengenai topic modeling telah berkembang pesat dalam beberapa tahun terakhir, khususnya pada analisis data media sosial yang memiliki karakteristik dinamis, tidak terstruktur, dan mengandung noise. Penelitian dari Blei et al. (2003), memperkenalkan *Latent Dirichlet Allocation* (LDA) sebagai pendekatan probabilistik untuk mengidentifikasi topik tersembunyi dalam kumpulan dokumen. Meskipun LDA menjadi salah satu metode yang paling banyak digunakan dalam text mining, beberapa penelitian menunjukkan bahwa metode ini memiliki keterbatasan dalam menangani data short text karena kurang mampu menangkap hubungan semantik dan konteks antar kata secara mendalam.

Seiring berkembangnya model transformer, penelitian dari Grootendorst (2022), mengembangkan BERTopic yang mengombinasikan *Sentence*-BERT, UMAP, HDBSCAN, dan *class-based* TF-IDF (c-TF-IDF) untuk menghasilkan representasi topik yang lebih kontekstual. Penelitian tersebut menunjukkan bahwa pendekatan berbasis embedding mampu menghasilkan topik yang lebih koheren dibandingkan metode probabilistik tradisional, terutama pada data teks pendek dan tidak terstruktur.

Pada studi terapan yang dilakukan oleh Nanayakkara dan Thennakoon (2024), melakukan perbandingan beberapa metode *topic modeling* pada data komentar media sosial dan menemukan bahwa BERTopic menghasilkan kualitas topik yang lebih baik dibandingkan LDA dan *Non-negative Matrix Factorization* (NMF) berdasarkan metrik *topic coherence*. Temuan serupa juga dilaporkan oleh Nugraha dan Utami (2024) yang membandingkan BERTopic dan LDA pada data ekonomi kreatif dan pariwisata, di

mana BERTopic menunjukkan kemampuan yang lebih baik dalam menangkap hubungan semantik antar dokumen.

Selain itu, Nursyahrina et al. (2024) menerapkan BERTopic dan LDA untuk menganalisis tren penelitian bidang ilmu komputer. Hasil penelitian menunjukkan bahwa BERTopic menghasilkan topik yang lebih koheren dan lebih mudah diinterpretasikan dibandingkan LDA. Meskipun demikian, penelitian tersebut masih menggunakan dokumen formal sehingga belum sepenuhnya merepresentasikan karakteristik data media sosial yang bersifat pendek, informal, dan mengandung noise tinggi.

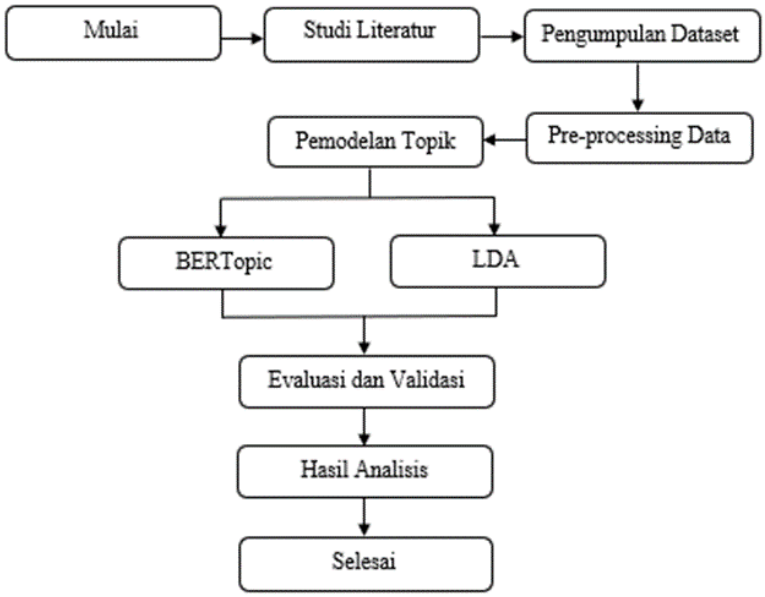
Berdasarkan kajian literatur tersebut, terdapat beberapa keterbatasan yang masih ditemukan pada penelitian sebelumnya. Pertama, sebagian besar penelitian menggunakan dataset berupa artikel ilmiah, berita, data pariwisata, atau media sosial umum, sehingga belum secara spesifik membahas fenomena promosi judi online. Kedua, penelitian terdahulu lebih banyak berfokus pada evaluasi performa topic modeling secara umum tanpa mengkaji kemampuan model dalam mengidentifikasi pola promosi konten ilegal pada platform media sosial. Ketiga, penelitian yang secara khusus membandingkan efektivitas LDA dan BERTopic pada komentar YouTube terkait promosi judi online masih sangat terbatas.

Berdasarkan keterbatasan tersebut, research gap yang diidentifikasi dalam penelitian ini adalah belum adanya kajian yang secara komprehensif mengevaluasi kemampuan metode probabilistik dan transformer-based topic modeling dalam menganalisis komentar YouTube yang berkaitan dengan promosi judi online. Karakteristik komentar YouTube yang pendek, tidak terstruktur, mengandung spam, bahasa informal, dan noise semantik menjadi tantangan yang berbeda dibandingkan dataset formal yang digunakan pada penelitian sebelumnya.

Oleh karena itu, penelitian ini memberikan kontribusi dalam tiga aspek utama. Pertama, penelitian ini menghadirkan studi komparatif antara LDA dan BERTopic pada domain promosi judi online yang masih relatif jarang diteliti dalam literatur *topic modeling*. Kedua, penelitian ini mengevaluasi kualitas topik menggunakan metrik *topic coherence* dan *topic diversity* sehingga dapat memberikan gambaran yang lebih komprehensif mengenai kualitas representasi topik yang dihasilkan. Ketiga, penelitian ini memberikan pemahaman mengenai efektivitas semantic topic modeling berbasis *transformer* dalam menganalisis data *short text* media sosial yang kompleks, serta mendukung pengembangan sistem monitoring konten ilegal berbasis *Natural Language Processing*.

3. Metodologi Penelitian

Pada tahapan ini membahas tentang penerapan perancangan metode terlebih dahulu untuk meminimalisir kesalahan dalam proses penelitian. Tahapan metodologi penelitian dapat dilihat pada Gambar 2.



Gambar 2 . Tahapan Metodologi Penelitian

3.1 Studi Literatur

Tahapan ini diawali dengan mempelajari berbagai literatur, baik berupa jurnal maupun buku yang berkaitan dengan topik penelitian. Kajian literatur difokuskan pada pembahasan mengenai metode *topic modeling*, khususnya *Latent Dirichlet Allocation* (LDA) dan BERTopic, serta tahapan preprocessing data teks. Selain itu, dilakukan kajian terhadap metrik evaluasi topik, yaitu *topic coherence* dan *topic diversity*, serta pemahaman mengenai penggunaan *library* seperti BERTopic dan OCTIS yang digunakan dalam implementasi dan evaluasi model.

3.2 Pengumpulan Dataset

Data yang digunakan dalam penelitian ini berupa kumpulan komentar terkait judi online pada platform YouTube yang diperoleh dari dataset publik di Kaggle (Yaemico, 2024) dengan total sebanyak 6.350 data. Dataset tersebut merupakan hasil pengumpulan data melalui proses *scraping* komentar YouTube pada konten yang relevan dengan topik judi online, khususnya pada siaran langsung (*live streaming*).

Data yang diperoleh telah melalui tahap validasi dan pembersihan (*data cleaning*), seperti penghapusan duplikasi, normalisasi teks, serta penyaringan komentar yang tidak relevan, sehingga data yang digunakan memiliki kualitas yang baik untuk dianalisis.

Dataset ini telah dilengkapi dengan label kelas, di mana label 0 menunjukkan komentar yang tidak mengandung promosi judi online, sedangkan label 1 menunjukkan komentar yang mengandung promosi judi online. Struktur data terdiri dari beberapa atribut, yaitu *author name, message, cleaned message*, dan label.

3.3 Pre-processing Data

Tahap preprocessing data dilakukan untuk menyiapkan data teks agar siap digunakan pada proses pemodelan topik. Pre-processing data bertujuan untuk membersihkan data dari kata atau karakter yang tidak diperlukan serta menyamakan bentuk teks, sehingga hasil analisis topik menjadi lebih jelas dan mudah dipahami.

a. Case Folding

Case Folding merupakan proses mengubah seluruh karakter pada teks menjadi huruf kecil dan huruf besar. Tahap ini dilakukan untuk menghindari perbedaan makna yang disebabkan oleh variasi penggunaan huruf kecil maupun huruf kapital dalam komentar YouTube. Hasil proses dapat dilihat pada Tabel 1.

Tabel 1. Hasil proses *Case Folding*

No.	***Message***	***Cleaned_message***
1.	info link gacor	info link gacor
2.	selamat ulang tahun Jogjaku tercinta, terimakasih sudah menjadi kota penuh kenangan	selamat ulang tahun jogjaku tercinta terimakasih sudah menjadi kota penuh kenangan
3.	assalamu'alaikum..	assalamualaikum
4.	MENDING NONTON LIVESTREAM SOALNYA JALAN KE JOGJA UDAH DI PENUH	mending nonton livestream soalnya jalan ke jogja udah di penuh

b. Tokenization

Tokenization merupakan proses memecah teks menjadi unit-unit kata atau token. Tahap ini bertujuan untuk memudahkan proses analisis teks dengan mengidentifikasi kata-kata yang membentuk setiap komentar. Hasil proses dapat dilihat pada Tabel 2.

Tabel 2. Hasil proses *Tokenization*

No.	***Cleaned_message***	***Tokenization***
1.	info link gacor	info,link,gacor
2.	selamat ulang tahun jogjaku tercinta terimakasih sudah menjadi kota penuh kenangan	selamat,ulang,jogjaku,tercinta, terimakasih,kota,penuh,kenangan
3.	assalamualaikum	assalamualaikum
4.	mending nonton livestream soalnya jalan ke jogja udah di penuh	mending,nonton,livestream ,jalan,jogja,udah,penuh

c. Stopword Removal

Stopword Removal dilakukan dengan menghapus kata-kata umum yang sering muncul namun tidak memiliki makna signifikan dalam pembentukan topik, seperti kata sambung dan kata ganti. Tahap ini bertujuan untuk meningkatkan relevansi kata-kata yang digunakan dalam pemodelan topik. Hasil proses dapat dilihat pada Tabel 3.

Tabel 3. Hasil proses stopword removal

No.	***Tokenization***	***Removed_Stopword***
1.	info,link,gacor	info link gacor
2.	selamat,ulang,jogjaku,tercinta, terimakasih,kota,penuh,kenangan	selamat ulang jogjaku tercinta terimakasih kota penuh kenangan
3.	assalamualaikum	assalamualaikum
4.	mending,nonton,livestram ,jalan,jogja,udah,penuh	mending nonton livestream jalan jogja udah di penuh

d. Stemming

Stemming merupakan proses mengubah kata ke bentuk dasarnya. Tahap ini dilakukan untuk menyatukan variasi kata yang memiliki makna yang sama sehingga dapat mengurangi dimensi data dan meningkatkan kualitas hasil pemodelan topik. Hasil proses dapat dilihat pada Tabel 4.

Tabel 4. Hasil proses stemming

No.	***Removed_Stopword***	***Stemming***
1.	info link gacor	info link gacor
2.	selamat ulang jogjaku tercinta terimakasih kota penuh kenangan	selamat ulang jogjaku cinta terimakasih kota penuh kenangan
3.	assalamualaikum	assalamualaikum
4.	mending nonton livestream jalan jogja udah di penuh	mending nonton livestream jalan jogja udah penuh

3.4 Pemodelan Topik

Pemodelan topik dilakukan untuk mengidentifikasi dan mengelompokkan topik-topik dominan dalam komentar YouTube terkait konten judi online. Pada penelitian ini digunakan dua metode, yaitu *Latent Dirichlet Allocation* (LDA) dan BERTopic. Metode LDA digunakan untuk memodelkan topik berdasarkan distribusi probabilitas kata dalam dokumen, sehingga dapat diketahui hubungan antara kata dan topik. Sementara itu, BERTopic digunakan untuk memodelkan topik berbasis representasi semantik dengan memanfaatkan embedding dari model transformer.

Hasil dari kedua metode tersebut kemudian dibandingkan untuk mengetahui metode yang lebih efektif dalam mengidentifikasi topik dominan pada data komentar YouTube.

3.5 Evaluasi dan Validasi

Evaluasi dan validasi dilakukan untuk menilai kualitas topik yang dihasilkan oleh metode *Latent Dirichlet Allocation* (LDA) dan BERTopic. Evaluasi pada kedua metode dilakukan menggunakan dua metrik utama, yaitu *topic coherence* dan *topic diversity*. *Topic coherence* digunakan untuk mengukur tingkat keterkaitan semantik antar kata dalam suatu topik, sedangkan *topic diversity* digunakan untuk mengukur tingkat keberagaman kata antar topik yang dihasilkan.

Secara matematis, *topic coherence* dihitung berdasarkan hubungan pasangan kata dalam topik, yang dirumuskan sebagai:

$$C_V = \frac{1}{T}\sum_{i=1}^{T} coherence(t_i) \quad \ldots.(3)$$

T merupakan jumlah topik, dan $coherence\,(t_i)$ skor koherensi pada topik ke-i yang merepresentasikan keterkaitan kemunculan bersama (co-occurrence) antar kata dalam dokumen. Pendekatan ini digunakan untuk menilai apakah kata-kata utama dalam suatu topik muncul secara konsisten dalam konteks yang sama (Handayani, 2026). Sementara itu, *topic diversity* dihitung dengan persamaan:

$$TD = \frac{|unique_{word}|}{|total_{word}|} \quad \ldots.(4)$$

Nilai *topic diversity* yang tinggi menunjukkan bahwa kata-kata antar topik semakin beragam, sedangkan nilai yang rendah mengindikasikan adanya kemiripan kata antar topik (Handayani, 2026).

Selain itu, evaluasi juga didukung dengan analisis interpretatif terhadap hasil topik untuk memastikan bahwa topik yang dihasilkan relevan dan dapat dipahami secara semantik. Proses validasi dilakukan dengan membandingkan hasil evaluasi dari kedua metode untuk menentukan metode yang paling optimal dalam mengidentifikasi topik dominan pada data komentar YouTube.

4. Hasil dan Pembahasan

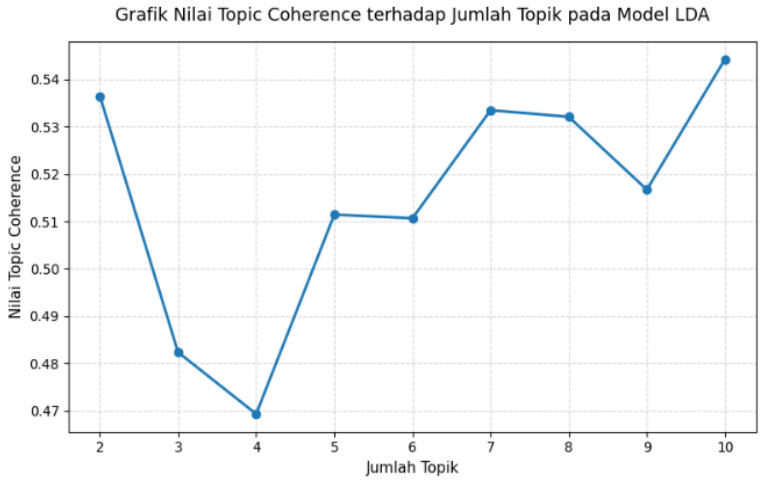
4.1. Pemodelan Topik menggunakan *Latent Dirichlet Allocation*

Pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) diterapkan pada data komentar YouTube yang berkaitan dengan aktivitas judi online untuk mengidentifikasi tema-tema utama yang muncul dalam data. Melalui pemodelan

tersebut, dapat diperoleh gambaran mengenai pola distribusi topik serta jumlah topik yang mudah dipahami dalam merepresentasikan keseluruhan komentar. Kualitas topik yang dihasilkan selanjutnya dievaluasi menggunakan metrik *topic coherence* sebagai dasar dalam proses analisis dan interpretasi topik.

a. Penentuan jumlah topik optimal pada model *Latent Dirichlet Allocation*

Penentuan jumlah topik pada model *Latent Dirichlet Allocation* (LDA) dilakukan dengan menguji variasi jumlah topik dari 2 hingga 10. Evaluasi dilakukan menggunakan metrik *topic coherence* (c_v) untuk mengukur keterkaitan semantik antar kata dalam setiap topik. Hasil pengujian nilai *coherence* untuk setiap jumlah topik dapat dilihat pada Gambar 3. Nilai *coherence* yang lebih tinggi menunjukkan bahwa topik yang dihasilkan memiliki konsistensi makna yang lebih baik antar kata dalam setiap topik.



Gambar 3. Grafik Nilai *Topic Coherence* terhadap Jumlah Topik (LDA)

Hasil pengujian menunjukkan bahwa nilai *topic coherence* berfluktuasi terhadap perubahan jumlah topik dan tidak menunjukkan pola linear. Nilai *coherence* tertinggi diperoleh pada jumlah topik 10 sebesar 0.544. Namun, model dengan 5 topik dipilih sebagai konfigurasi optimal dengan nilai *coherence* sebesar 0.511. Pemilihan ini didasarkan pada pertimbangan keseimbangan antara kualitas model dan interpretabilitas topik. Jumlah topik yang terlalu besar cenderung menghasilkan topik yang lebih spesifik namun terpisah-pisah sehingga sulit diinterpretasikan. Oleh karena itu, penggunaan 5 topik dinilai lebih representatif dalam menggambarkan struktur tema pada data komentar.

b. Evaluasi model *Latent Dirichlet Allocation*

Evaluasi model *Latent Dirichlet Allocation* (LDA) dilakukan menggunakan dua metrik, yaitu *topic coherence* dan *topic diversity*. *Topic coherence* digunakan untuk

mengukur tingkat keterkaitan makna antar kata dalam suatu topik. Pada model LDA dengan jumlah 5 topik sebagai model final, diperoleh nilai *topic coherence* sebesar 0.511. Nilai ini menunjukkan bahwa kata-kata penyusun topik memiliki keterkaitan semantik yang cukup baik, sehingga topik yang dihasilkan dapat diinterpretasikan secara jelas.

Nilai *topic coherence* tersebut mengindikasikan bahwa model LDA mampu menangkap pola distribusi kata yang relevan pada data komentar YouTube terkait judi online. Meskipun nilai yang diperoleh tidak merupakan yang tertinggi dibandingkan variasi jumlah topik lainnya, kualitas topik yang dihasilkan tetap tergolong memadai untuk mendukung analisis tematik.

Selain itu, evaluasi juga dilakukan menggunakan metrik *topic diversity* untuk mengukur tingkat keberagaman kata antar topik. Hasil perhitungan menunjukkan bahwa model LDA menghasilkan nilai *topic diversity* sebesar 1.0. Nilai ini mengindikasikan bahwa tidak terdapat tumpang tindih kata antar topik, sehingga setiap topik memiliki karakteristik yang berbeda secara jelas. Namun, nilai *topic diversity* yang tinggi juga dapat menunjukkan rendahnya keterkaitan antar topik, karena tidak adanya kata yang digunakan secara bersama.

Secara keseluruhan, berdasarkan nilai *topic coherence* dan *topic diversity*, model LDA mampu menghasilkan topik yang cukup koheren dengan tingkat keberagaman yang tinggi. Dengan demikian, metode LDA dapat digunakan sebagai pendekatan awal dalam pemodelan topik berbasis distribusi kata untuk menggambarkan tema-tema utama pada data komentar YouTube. Hasil ini selanjutnya dibandingkan dengan metode BERTopic untuk mengevaluasi performa kedua metode dalam menghasilkan topik yang lebih optimal.

c. Hasil pembentukan dan interpretasi topik *Latent Dirichlet Allocation*

Proses pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA) menghasilkan 5 topik utama yang dipilih sebagai model terbaik berdasarkan nilai *topic coherence*. Setiap topik direpresentasikan oleh sekumpulan kata dengan probabilitas tertinggi yang mencerminkan tema tertentu dalam data komentar YouTube terkait judi online.

Tabel 5. Hasil Pemodelan Topik Menggunakan LDA

Topic	Kata Representatif dan Probabilitas
0	0.102*"jogja" + 0.048*"istimewa" + 0.033*"selamat" + 0.032*"kota" + 0.023*"ulang" + 0.012*"yogyakarta" + 0.012*"hbd" + 0.010*"drone" + 0.009*"iklan" + 0.009*"ku"

1	0.020*"nonton" + 0.015*"ambal" + 0.014*"ga" + 0.013*"sugeng" + 0.012*"yg" + 0.011*"warsa" + 0.009*"tugu" + 0.008*"mantap" + 0.007*"selesai" + 0.007*"bgt"
2	0.147*"rb" + 0.121*"firewisdomtotofire" + 0.117*"freebet" + 0.116*"wisdomtotofire" + 0.094*"ketik" + 0.057*"googlewisdomtoto" + 0.051*"slot" + 0.044*"gacorfire" + 0.037*"google" + 0.036*"fire"
3	0.033*"ning" + 0.027*"yaaa" + 0.018*"judol" + 0.016*"nyaman" + 0.015*"nong" + 0.015*"gung" + 0.015*"bayanjogja" + 0.015*"tetaplah" + 0.013*"tolong" + 0.013*"pulang"
4	0.106*"fireangka" + 0.106*"terbalik" + 0.106*"bayarfire" + 0.017*"menyala" + 0.012*"hadir" + 0.007*"kembang" + 0.007*"lho" + 0.006*"gacor" + 0.006*"brp" + 0.006*"gacorr"

Berdasarkan tabel 5 hasil pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA), diperoleh lima topik utama yang merepresentasikan pola kata dalam data komentar YouTube terkait judi online. Setiap topik ditunjukkan oleh kumpulan kata dengan probabilitas tertinggi yang mencerminkan tema tertentu.

Topik 0 didominasi oleh kata-kata seperti jogja, istimewa, selamat, dan kota yang menunjukkan konteks komentar umum yang tidak secara langsung berkaitan dengan judi online. Topik ini kemungkinan merepresentasikan komentar yang bersifat umum atau tidak relevan dengan fokus penelitian.

Topik 1 mengandung kata-kata seperti nonton, ga, sugeng, dan mantap yang menunjukkan bentuk komentar interaksi pengguna yang bersifat umum, seperti respon terhadap video. Topik ini juga tidak memiliki keterkaitan langsung dengan aktivitas judi online.

Topik 2 merupakan topik yang paling relevan dengan konteks penelitian karena mengandung kata-kata seperti freebet, slot, google, dan fire yang berkaitan dengan aktivitas serta promosi judi online. Kata-kata tersebut menunjukkan adanya indikasi promosi atau penyebaran informasi terkait situs judi online dalam komentar YouTube.

Topik 3 terdiri dari kata-kata seperti ning, yaaa, judol, dan nyaman yang menunjukkan kombinasi antara kata tidak baku dan istilah terkait judi online (judol). Namun, topik ini masih mengandung banyak kata yang kurang jelas sehingga sulit diinterpretasikan secara spesifik, yang kemungkinan disebabkan oleh adanya noise pada data.

Topik 4 mengandung kata-kata seperti fireangka, bayarfire, gacor, dan menyala yang juga memiliki keterkaitan dengan aktivitas judi online, khususnya dalam konteks promosi dan istilah yang sering digunakan dalam perjudian.

Meskipun demikian, masih terdapat beberapa kata yang kurang representatif sehingga interpretasi topik ini tidak sepenuhnya jelas.

Secara keseluruhan, hasil pemodelan LDA menunjukkan bahwa tidak semua topik yang dihasilkan memiliki keterkaitan langsung dengan judi online. Hal ini disebabkan oleh karakteristik data komentar YouTube yang bersifat tidak terstruktur, mengandung bahasa tidak baku, serta adanya noise dalam data. Namun demikian, beberapa topik yang terbentuk, khususnya Topik 2 dan Topik 4, berhasil mengidentifikasi pola kata yang berkaitan dengan promosi judi online.

4.2. Pemodelan Topik menggunakan BERTopic

Pemodelan topik menggunakan BERTopic dilakukan untuk mengidentifikasi tema-tema utama dalam komentar YouTube terkait judi online dengan memanfaatkan representasi semantik berbasis embedding. Berbeda dengan LDA yang mengandalkan distribusi kata, BERTopic mengelompokkan dokumen berdasarkan kemiripan makna kalimat, sehingga diharapkan mampu menghasilkan topik yang lebih koheren secara semantik.

a. Hasil pembentukan topik menggunakan BERTopic

Pemodelan topik menggunakan BERTopic dilakukan dengan memanfaatkan representasi semantik berbasis embedding untuk mengelompokkan komentar YouTube yang berkaitan dengan judi online berdasarkan kemiripan makna. Proses ini memungkinkan terbentuknya topik-topik yang tidak hanya bergantung pada kemunculan kata, tetapi juga pada konteks kalimat secara keseluruhan.

Tabel 6. Hasil Pemodelan Topik Menggunakan BERTopic

ID Topik	Kata Kunci Representatif	Interpretasi	Jumlah Data
-1	Kata umum, tidak spesifik, percakapan acak	Outlier	1210
0	Jogja, madang, makan, pesan, kuliner	Aktivitas kuliner/ percakapan umum	192
1	Bayarfire, freebet, slot, gacor, menang	Promosi judi online	98
2	Aku, males, telat, jalan, pulang	Percakapan pribadi/aktivitas sehari-hari	95
3	Freebet, slot, bonus, akun, daftar	Strategi pemasaran judi	75

Berdasarkan tabel 6. pemodelan menggunakan metode BERTopic, diperoleh beberapa topik yang terbentuk berdasarkan kesamaan konteks semantik dari komentar YouTube. Setiap topik direpresentasikan oleh kata

kunci serta contoh dokumen yang menggambarkan isi topik secara lebih kontekstual.

Topik dengan jumlah data terbesar (*Topic* -1) menunjukkan kumpulan komentar yang tidak dapat dikelompokkan secara spesifik (outlier), yang umumnya disebabkan oleh variasi bahasa dan ketidakteraturan teks.

Selain itu, topik lain yang terbentuk seperti *Topic* 0 menunjukkan kelompok komentar dengan tema tertentu yang lebih terstruktur, ditandai dengan kemunculan kata-kata yang saling berkaitan dalam satu konteks. Hal ini menunjukkan bahwa BERTopic mampu mengelompokkan komentar berdasarkan makna kalimat, bukan hanya frekuensi kata.

Secara keseluruhan, hasil BERTopic menunjukkan kemampuan dalam menangkap konteks semantik yang lebih baik dibandingkan metode berbasis distribusi kata, meskipun masih terdapat outlier yang cukup besar akibat karakteristik data komentar yang tidak terstruktur.

b. Evaluasi model BERTopic

Evaluasi kualitas topik pada model BERTopic dilakukan menggunakan metrik *topic coherence* dan *topic diversity*. *Topic coherence* digunakan untuk mengukur keterkaitan makna antar kata dalam satu topik, sedangkan *topic diversity* digunakan untuk menilai keberagaman kata antar topik.

Berdasarkan hasil perhitungan, model BERTopic menghasilkan nilai *topic coherence* sebesar 0.667. Nilai ini menunjukkan bahwa kata kunci dalam setiap topik memiliki keterkaitan makna yang cukup baik, sehingga topik yang dihasilkan relatif mudah untuk diinterpretasikan. Hal ini mengindikasikan bahwa pendekatan berbasis embedding yang digunakan oleh model mampu menangkap kesamaan konteks antar komentar secara efektif.

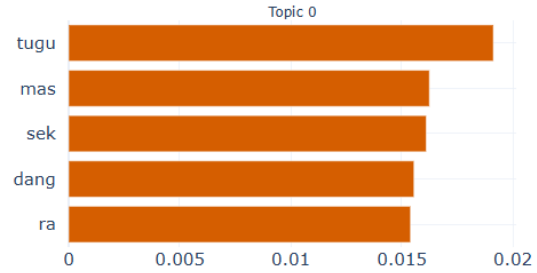
Selain itu, model BERTopic menghasilkan nilai *topic diversity* sebesar 0.449. Nilai ini menunjukkan bahwa terdapat beberapa kata yang muncul pada lebih dari satu topik, sehingga masih terdapat tingkat overlap antar topik. Hal ini mengindikasikan adanya kemiripan konteks antar topik yang dihasilkan, sehingga keberagaman kata menjadi tidak terlalu tinggi.

Secara keseluruhan, model BERTopic mampu menghasilkan topik yang cukup representatif terhadap variasi komentar yang terdapat dalam data, seperti komentar promosi judi online, spam, maupun tanggapan pengguna, meskipun masih terdapat beberapa kemiripan antar topik.

c. Visualisasi dan Interpretasi Topik menggunakan BERTopic

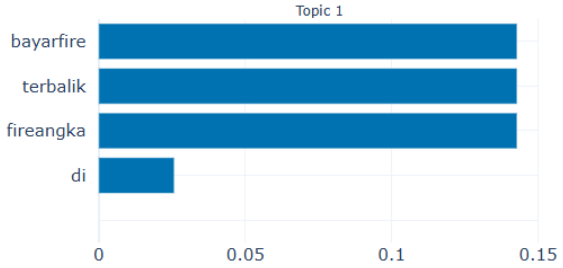
Visualisasi hasil pemodelan topik menggunakan BERTopic ditampilkan dalam bentuk *bar chart topic word scores*. *Topic word scores* merupakan nilai bobot kata yang dihitung menggunakan metode *class-based* TF-IDF (c-TF-IDF), yang

menunjukkan tingkat kepentingan suatu kata dalam merepresentasikan topik tertentu (Wahyuni et al., 2025).



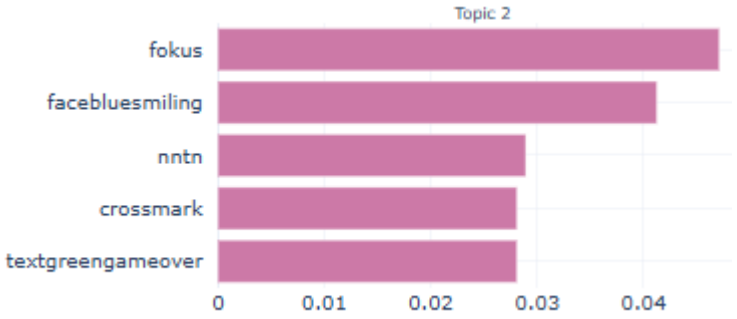
Gambar 4. Topic 0

Gambar 4 menunjukkan *Topic* 0 yang didominasi oleh kata-kata seperti tugu, mas, sek, dang, dan ra. Kata-kata tersebut tidak membentuk tema yang jelas dan cenderung merepresentasikan percakapan umum.



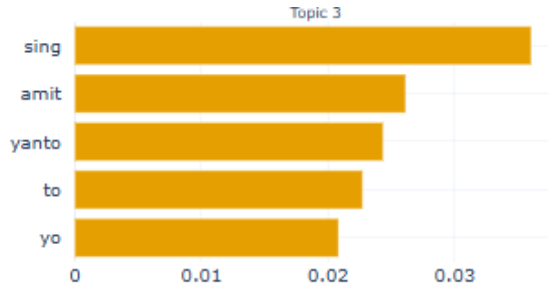
Gambar 5. Topic 1

Gambar 5 menunjukkan topik 1 yang didominasi oleh kata bayarfire, terbalik, dan fireangka. Kata-kata ini berkaitan dengan perjudian online, sehingga topik ini dapat diinterpretasikan sebagai aktivitas atau promosi judi online.



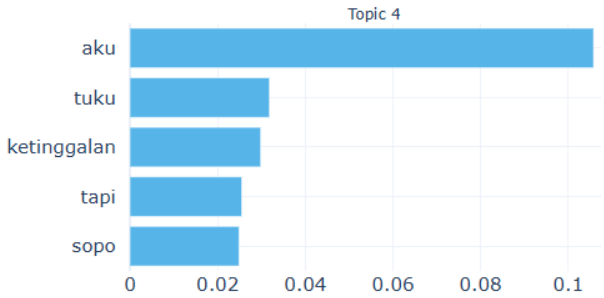
Gambar 6. Topic 2

Gambar 6 menunjukkan Topik 2 yang terdiri dari kata seperti fokus, facebluesmiling, ntnn, dan crossmark. Kata tersebut menggambarkan ekspresi atau reaksi pengguna, sehingga topik ini merepresentasikan interaksi atau respon pengguna terhadap konten.



Gambar 7. Topic 3

Gambar 7 menunjukkan Topic 3 yang didominasi oleh kata seperti sing, amit, yanto, to, dan yo. Kata-kata ini cenderung merupakan potongan kata atau bahasa informal, sehingga topik ini tidak memiliki makna yang jelas dan termasuk dalam kategori percakapan umum.



Gambar 8. Topic 4

Gambar 8 menunjukkan Topic 4 yang didominasi oleh kata seperti aku, tuku, ketinggalan, tapi, dan sopo. Kata-kata tersebut menggambarkan percakapan pribadi pengguna, sehingga topik ini merepresentasikan aktivitas sehari-hari atau opini individu.

Secara keseluruhan, visualisasi menunjukkan bahwa BERTopic mampu mengelompokkan data berdasarkan konteks semantik, di mana beberapa topik memiliki keterkaitan dengan promosi judi online, sementara topik lainnya dipengaruhi oleh noise pada data komentar.

5. Kesimpulan

Penelitian ini melakukan evaluasi komparatif antara *Latent Dirichlet Allocation* (LDA) dan BERTopic untuk menganalisis komentar YouTube yang berkaitan dengan promosi

judi online. Hasil penelitian menunjukkan bahwa BERTopic menghasilkan nilai *topic coherence* sebesar 0.667, lebih tinggi dibandingkan LDA sebesar 0.511, yang menunjukkan kemampuan lebih baik dalam menghasilkan topik yang koheren secara semantik. Sebaliknya, LDA memperoleh nilai *topic diversity* sebesar 1.0, lebih tinggi dibandingkan BERTopic sebesar 0.449, yang mengindikasikan keberagaman topik yang lebih luas.

Temuan ini menunjukkan bahwa pendekatan transformer-based topic modeling lebih efektif dalam menangkap hubungan kontekstual dan representasi semantik pada data komentar media sosial yang bersifat pendek, tidak terstruktur, dan mengandung noise. Hasil penelitian ini memperkuat perkembangan paradigma semantic topic modeling dalam bidang *Natural Language Processing,* khususnya pada analisis *short-text social media* data. Dengan demikian, BERTopic dapat menjadi alternatif yang lebih sesuai dibandingkan metode probabilistik tradisional untuk analisis diskursus digital yang kompleks.

Penelitian selanjutnya diharapkan dapat memanfaatkan dataset yang lebih beragam dari berbagai platform media sosial serta mengombinasikan metode *topic modeling* dengan pendekatan lain, seperti *social network analysis* atau *deep learning*, guna meningkatkan kualitas pemodelan topik dan akurasi dalam mendeteksi konten judi online.

Secara praktis, hasil penelitian ini dapat dimanfaatkan dalam pengembangan sistem monitoring konten digital, deteksi promosi judi online, serta pengawasan konten ilegal pada platform media sosial. Selain itu, temuan penelitian ini berpotensi mendukung pengambilan kebijakan berbasis data dalam upaya mitigasi penyebaran konten perjudian daring di ruang digital.

Penelitian ini masih memiliki beberapa keterbatasan, antara lain penggunaan satu sumber dataset, fokus pada komentar YouTube berbahasa Indonesia, serta penggunaan metrik evaluasi yang terbatas pada *topic coherence* dan *topic diversity*. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset multibahasa, membandingkan metode topic modeling yang lebih beragam seperti Top2Vec dan *Contextualized Topic Model* (CTM), serta mengintegrasikan evaluasi berbasis pakar untuk meningkatkan validitas interpretasi topik yang dihasilkan.

Secara keseluruhan, penelitian ini menunjukkan bahwa pemodelan topik berbasis transformer memiliki potensi yang signifikan dalam mendukung analisis konten media sosial yang kompleks dan dapat menjadi fondasi bagi pengembangan sistem deteksi konten ilegal berbasis *Artificial Intelligence* di masa depan.

6. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Universitas Multi Data Palembang, khususnya dosen pembimbing, atas dukungan dan bimbingan selama proses penelitian.

Ucapan terima kasih juga disampaikan kepada penyedia dataset melalui platform Kaggle atas ketersediaan data yang digunakan dalam penelitian ini.

7. Pernyataan Penulis

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menyatakan bahwa data dan makalah bebas dari plagiarisme serta penulis bertanggung jawab secara penuh atas keaslian artikel.

Bibliografi

Blei, D. M., Ng, A. Y., & Jordan, M. T. (2003). Latent dirichlet allocation. Advances in Neural Information Processing Systems, 3, 993–1022.

David, E., Sondakh, M., & Harilama, S. (2017). Pengaruh Konten Vlog dalam Youtube terhadap Pembentukan Sikap Mahasiswa Ilmu Komunikasi. Acta Diurna, 6(1). https://ejournal.unsrat.ac.id/index.php/index/index

Faizah, & Lin, B. S. (2023). Visualizing Change and Correlation of Topics With LDA and Agglomerative Clustering on COVID-19 Vaccine Tweets. IEEE Access, 11(June), 51647–51656. https://doi.org/10.1109/ACCESS.2023.3278979

Gunadi, I. M. D. A., & Sugiantari, A. A. W. (2024). Mekanisme dan regulasi penegakan hukum terhadap streamer game yang menyampaikan informasi tentang judi online di YouTube. Jurnal Hukum Mahasiswa, 4(1). https://doi.org/10.36733/jhm.v4i1

Grehenson, G. (2024). Judi Online Makin Marak di Kalangan Anak Muda, Pakar UGM Sarankan Perlunya Edukasi Literasi Keuangan. UNIVERSITAS GADJAH MADA. https://ugm.ac.id/id/berita/judi-online-makin-marak-di-kalangan-anak-muda-pakar-ugm-sarankan-perlunya-edukasi-literasi-keuangan/

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794

Handayani, L. N. (2026). Analisis perbandingan performa NMF dengan LDA pada topik modeling berita online Indonesia (Tesis Magister). Universitas Teknologi Digital Indonesia.

Husain, W. R. A.-F. (2024). Hukum Pidana Judi Online Perspektif Indonesia Dan Perkembangan Aspek Legalitas. Journal Of Human And Education (JAHE), 4(6), 1297–1304. https://doi.org/10.31004/jh.v4i6.2049

Indra, S. M., & Srihadiati, T. (2025). Analisis kriminologi peran konstruksi media terhadap penyebaran konten judi online dalam media sosial Facebook. Ranah Research: Journal of Multidisciplinary Research and Development, 7(5)

Irawan, H. (2024). Regulasi hukum bisnis dalam praktik endorsement judi online di media sosial oleh influencer Indonesia: A review. Islamic Law Journal, 2(2).

Jelita, M. (2024). Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI. Media Pustakawan, 31(3).

Kannitha, D. Z. T., Mustafid, M., & Kartikasari, P. (2022). Pemodelan Topik Pada Keluhan Pelanggan Menggunakan Algoritma Latent Dirichlet Allocation Dalam Media Sosial Twitter. Jurnal Gaussian, 11(2), 266–277. https://doi.org/10.14710/j.gauss.v11i2.35474

Nanayakkara, A. C., & Thennakoon, G. A. D. M. (2024). Enhancing Social Media Content Analysis with Advanced Topic Modeling Techniques: A Comparative Study. International Journal on Advances in ICT for Emerging Regions (ICTer), 17(1), 40–47. https://doi.org/10.4038/icter.v17i1.7276

Nura Nugraha, I., & Utami, E. (2024). Evaluation of Creative Economy and Tourism Industry Trends based on LDA Analysis with BERTopic. Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi, 15(2), 182–195. https://doi.org/10.31849/digitalzone.v15i2.23796

Nursyahrina, N., Sarjon Defit, & Rini Sovia. (2024). Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer. Jurnal KomtekInfo, 11(4). DOI:10.35134/komtekinfo.v11i4.580

Pusat Pelaporan dan Analisis Transaksi Keuangan. (2026). Catatan capaian strategis PPATK tahun 2025: Menjaga kedaulatan dan integritas ekonomi bangsa. https://www.ppatk.go.id/siaran_pers/read/1594/

Samuel, & Kristiadi, D. P. (2024). Deteksi Teks Promosi Judi Online Menggunakan AI dengan Kombinasi NLP dan Deep Learning. Jurnal Sistem Informasi dan Teknologi (SINTEK), 5(2).

Sri Gustina, Alfarel Kurniawan, & Yusril Pandawa. (2025). Tindak Pidana Judi Online : Penegakan Hukum Oleh Kepolisian, Serta UpayaDan Strategi PenanganannyaOnline Gambling Crime: Law Enforcement by the Police, as well as Efforts andStrategies for Handling it. Jiic: Jurnal Intelek Insan Cendikia, 2(5), 7763–7776. https://jicnusantara.com/index.php/jiic

Syaifuddin, A., Harianto, R. A., & Santoso, J. (2021). Analisis Trending Topik untuk Percakapan Media Sosial dengan Menggunakan Topic Modelling Berbasis Algoritme LDA. Journal of Intelligent System and Computation, 2(1), 12–19. https://doi.org/10.52985/insyst.v2i1.150

Vigar, L. S., Himawan, K. K., & Mutiara, E. (2019). Hubungan antara Spiritualitas dan Religiusitas dengan Illusion of Control pada Emerging Adult. In Jurnal Ilmiah Psikologi MIND SET (Vol. 7, Issue 01, pp. 17–24). https://doi.org/10.35814/mindset.v7i01.305

Wahyuni, W., Lestari, T. P., Apriliana, M., & Gumelta, R. (2025). Implementation of BERTopic for topic modeling analysis of the free nutritious meal program based

on YouTube comments. Journal of Applied Informatics and Computing, 9(4), 1964–1971.

Syahindra, W., Murlena, M., & Hartati, H. (2020). Pemodelan Implementasi Open Access Repository Menggunakan Eprints Software di IAIN Curup. *Khizanah Al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, Dan Kearsipan*, *8*(1), 56–70. https://doi.org/10.24252/kah.v8i1a6

Yaemico. (2024). Deteksi Judi Online [Dataset]. Kaggle. https://www.kaggle.com/datasets/yaemico/deteksi-judi-online