Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 3, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i3.8763

Pemodelan Topik pada Komentar YouTube Arra: Komparasi LDA dan *K-Means* Menggunakan Fitur Leksikal dan Semantik

Siti Nuradilla[1], Sabrina Adnin Kamila[2], Latifah Zahra[3], Cici Suhaeni[4], Bagus Sartono[5]
[1,2,3,4,5]Department of Statistics, IPB University, Kampus Dramaga Bogor, 16680, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-05-15
Revised 2025-07-16
Accepted 2025-07-17

Abstract – *YouTube has become a platform for sharing content, including positive material and stereotypes that often trigger debates. One noteworthy phenomenon is the video of Arra, a toddler known for her remarkable communication skills. This uniqueness has drawn significant attention and sparked debates about the mismatch between her age and cognitive development. The diverse comments on Arra's videos reflect sharply differing perspectives among netizens, making manual analysis highly challenging. Therefore, it is important to examine the topics discussed by netizens to understand the dominant issues emerging in these discussions. Through this approach, the public can gain insights, and parents may receive valuable input regarding child-rearing practices. The main objective of this study is to explore the effectiveness of the two methods and their combinations of text representations in identifying key topics within comments by comparing the coherence performance of the models. This research applies topic modeling to analyze comments using two primary approaches: Latent Dirichlet Allocation (LDA) and K-Means clustering. The study involves data collection through comment crawling, followed by text preprocessing and text representation using TF-IDF and GloVe embeddings. LDA and K-Means are then used to identify dominant topics appearing in the comments. The results show that LDA with TF-IDF achieved the highest coherence score of 0.662, although the resulting topics were still difficult to interpret due to overlap. Meanwhile, K-Means with GloVe 100D yielded a slightly lower coherence score of 0.6538 but outperformed in terms of interpretability. Therefore, K-Means with GloVe 100D is considered a more balanced approach in terms of both coherence and topic readability.*

Keywords: *GloVe, K-Means, LDA, TF-IDF, Topic Modeling*

Corresponding Author:

Siti Nuradilla

Email: sitinuradilla@apps.ipb.ac.id

Abstrak – *YouTube menjadi platform untuk menyebarkan konten, baik konten positif maupun stereotip yang sering memicu perdebatan. Salah satu fenomena yang menarik perhatian adalah video Arra, seorang balita yang dikenal karena kemampuan komunikasinya yang bagus. Keunikan ini membuatnya menjadi sorotan, namun juga memicu perdebatan mengenai ketidaksesuaian usianya dengan pola pikirnya. Komentar-komentar yang beragam di video Arra mencerminkan perbedaan perspektif yang tajam di kalangan netizen, yang membuat analisis manual menjadi sangat menantang. Oleh karena itu, penting untuk mengecek pembahasan topik yang diungkapkan netizen, agar dapat memahami isu-isu dominan yang berkembang dalam diskusi tersebut. Utamanya, melalui pendekatan ini, masyarakat dapat mengambil pelajaran dan orang tua sendiri mendapatkan masukan terkait pola pendidikan pada anak. Tujuan utama penelitian ini yaitu untuk mengeksplorasi efektivitas kedua metode dan kombinasi repesentasi teks dalam mengidentifikasi topik-topik utama pada komentar melalui komparasi performa koherensi dari model. Penelitian ini menerapkan metode topic modeling untuk menganalisis komentar menggunakan dua pendekatan utama: Latent Dirichlet Allocation (LDA) dan K-Means. Penelitian ini mencakup pengumpulan data melalui crawling komentar dilanjutkan dengan proses pra-pemrosesan teks dan representasi teks menggunakan TF-IDF dan GloVe. LDA dan K-Means digunakan untuk mengidentifikasi topik-topik dominan yang muncul dalam komentar. Hasil menunjukkan bahwa LDA dengan TF-IDF menghasilkan coherence score tertinggi sebesar 0.662, namun topik-topik yang dihasilkan masih sulit untuk diinterpretasikan karena tumpeng tindih. Sementara, K-Means dengan GloVe 100D memiliki coherence score sedikit lebih rendah yakni 0.6538 namun lebih unggul dalam hal interpretabilitas. Dengan demikian, K-Means dengan Glove 100D menjadi pendekatan yang lebih seimbang antara koherensi dan keterbacaan topik.*

Kata Kunci*: GloVe, K-Means, LDA, TF-IDF, Topic Modeling*

I. PENDAHULUAN

Media sosial, terutama YouTube, menjadi *platform* yang efektif untuk menyebarkan berbagai konten yang sering memicu perdebatan dan kontroversi. Salah satu fenomena viral yang menarik perhatian adalah video Arra, seorang balita yang dikenal karena kemampuan komunikasinya yang luar biasa. Di usia yang masih muda, Arra mampu berdialog dengan kedua orang tuanya, mengenai topik-topik berat seperti pandangan hidup [1]. Keunikan ini membuat Arra menjadi idola baru di dunia maya, meskipun videonya sering kali memicu perdebatan, baik yang mendukung maupun yang menentang. Perdebatan ini membahas mengenai apakah perilaku dan pandangan Arra sesuai dengan usianya yang balita.

Beragam opini yang muncul pada komentar YouTube Arra, menunjukkan adanya perbedaan pandangan yang tajam di kalangan netizen. Jika dilakukan secara manual, sulit untuk menganalisis bahan perdebatan dan pandangan netizen karena jumlah komentar yang sangat banyak dan keragaman perspektif yang ada. Oleh karena itu, *topic modeling* menjadi sangat relevan untuk digunakan karena metode ini mampu mengidentifikasi topik-topik yang dominan pada komentar-komentar tersebut secara otomatis dan komprehensif [2]. Melalui pemanfaatan *topic modeling*, dapat diketahui topik-topik yang paling banyak dibahas oleh masyarakat.

Urgensi penggunaan *topic modeling* yakni penting untuk memahami bagaimana publik merespons video Arra dan bagaimana fenomena ini memengaruhi pandangan mereka. Melalui informasi mengenai topik-topik yang sering dibahas, dapat digali lebih dalam mengenai opini publik, baik yang mendukung maupun yang menentang pola asuh Arra. Informasi ini dapat membantu masyarakat untuk lebih bijak dalam mengasuh anak, serta memberikan masukan yang lebih konstruktif bagi orang tua, khususnya orang tua Arra. Salah satu pendekatan dalam *topic modeling* yang dapat digunakan adalah *Latent Dirichlet Allocation* (LDA) dan *K-Means* yang efektif mengidentifikasi pembahasan utama dalam komentar-komentar tersebut.

LDA merupakan model probabilistik generatif yang digunakan dalam analisis teks. Prinsip dasar LDA yakni dokumen-dokumen dalam korpus dianggap sebagai kombinasi dari berbagai topik tersembunyi, di mana setiap topik diwakili oleh distribusi probabilitas kata-kata tertentu [3]. Gambarannya, pada setiap komentar, akan terdapat lebih dari satu topik, topik yang paling dominan dilihat dari nilai probabilitasnya yang paling tinggi. Pada LDA, topik-topik didefinisikan berdasarkan probabilitas kata yang muncul dalam setiap topik, dan kata-kata dengan probabilitas tertinggi dalam suatu topik memberikan gambaran yang jelas mengenai makna atau fokus dari topik tersebut. Meskipun LDA efektif dalam menemukan topik, terdapat beberapa kekurangan yang dapat mempengaruhi kualitas hasilnya. LDA sering kali menghasilkan topik yang susah diinterpretasikan, terutama ketika topik-topik tersebut tumpang tindih, serta tidak efektif dalam menangani hubungan *non-linear* yang kompleks antara kata-kata dalam dokumen [4]. LDA juga umumnya tidak melakukan *word embedding* yang dapat memaksimalkan pemahaman semantik dan mengurangi *sparse* sehingga model kurang representatif.

Berbeda dari LDA, *K-Means* menjadi salah satu metode *clustering* yang dapat digunakan dalam memodelkan topik yang tidak bergantung pada distribusi probabilistik. Algoritma ini bekerja dengan mengelompokkan dokumen ke dalam sejumlah *cluster* berdasarkan kemiripan bobot dan vektornya, yang dihitung menggunakan representasi teks [5]. Beberapa teknik representasi teks yang sering digunakan yaitu TF-IDF dan *GloVe*. Penggunaan TF-IDF memungkinkan pemberian bobot yang lebih relevan pada kata-kata yang memiliki signifikansi tinggi dalam konteks dokumen [6]. *GloVe* sebagai model berbasis vektor kata memanfaatkan statistik global dari seluruh korpus untuk menangkap hubungan semantik antar kata yang lebih baik [7]. Melalui teknik representasi teks ke bentuk vektor, *K-Means* dapat fokus pada kata-kata yang lebih penting, mengurangi efek dari kata-kata yang jarang muncul, dan meningkatkan kualitas *clustering*. Berkaitan dengan hal ini, salah satu keunggulan *K-Means* adalah kemampuannya untuk menangani masalah *sparsity* yang sering terjadi ketika banyak kata hanya muncul sekali atau dua kali dalam dataset [8].

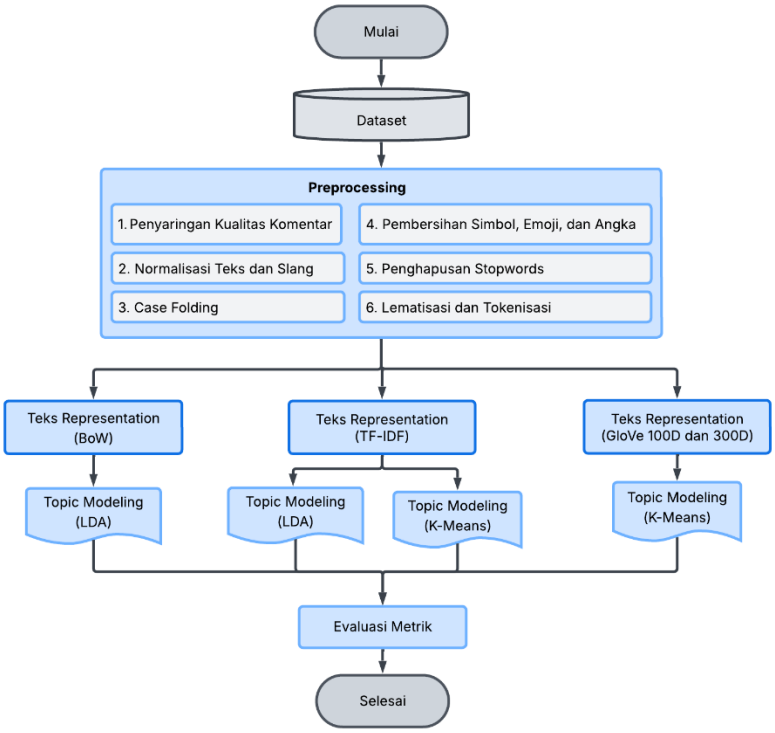
Penelitian yang membahas pemodelan LDA menggunakan TF-IDF sudah dilakukan pada topik pariwisata yang dengan nilai koherensi 0,613 dan 0,528 [9]. Sayangnya penelitian ini tidak secara langsung membandingkan dengan LDA model dasar, yaitu penggunaan *Bag of Words* (BoW) yang tidak melihat bobot kepentingan kata. Penelitian lain juga telah melakukan pemodelan topik menggunakan *K-Means* dengan representasi teks menggunakan *N-Grams* [10]. Penelitian tersebut masih berdasarkan representasi teks dengan fitur leksikal dan tidak mengkaji secara semantik pada konteks kalimat. Hal ini tentu tidak bisa dikaji dan dianalisis apakah penggunaan representasi teks berbasis semantik mampu menghasilkan pemodelan topik yang bagus.

Menyikapi problematika yang telah dipaparkan di atas, penelitian ini akan mengeksplorasi pemodelan topik pada komentar YouTube Arra dengan mengadopsi pendekatan berbasis analisis topik yang berbeda. Pada penelitian ini, digunakan metode LDA dan *K-Means* untuk memodelkan topik, namun dengan representasi teks yang membandingkan fitur leksikal dan semantik. Adapun teknik representasi teks sebagai pra-pemrosesan data yaitu TF-IDF dan *GloVe*. Selanjutnya, penelitian ini akan membandingkan hasil koherensi dan performa dari berbagai pendekatan yang menggabungkan teknik-teknik tersebut. Meskipun metode LDA dan K-Means telah banyak digunakan dalam pemodelan topik, studi yang secara eksplisit membandingkan keduanya dalam konteks komentar media sosial menggunakan dua pendekatan representasi teks yang berbeda, yakni leksikal (TF-IDF) dan semantik (GloVe) masih terbatas.

Penelitian ini menawarkan kebaruan dengan menghadirkan evaluasi sistematis terhadap kombinasi metode pemodelan topik dan representasi teks tersebut dalam konteks data yang bersifat informal, singkat, dan sangat beragam seperti komentar YouTube. Tujuan dari penelitian ini yaitu untuk mengeksplorasi keefektifan berbagai kombinasi model (LDA dan *K-Means*) dengan teknik representasi teks dalam mengidentifikasi topik-topik dominan pada komentar YouTube Arra. Hasil yang diperoleh diharapkan dapat memberikan wawasan tentang bagaimana metode-metode ini bekerja dalam mengidentifikasi isu-isu dominan, serta mengoptimalkan pemilihan teknik yang efektif dalam menangani komentar dalam jumlah besar dan beragam.

II. METODE

Metodologi yang digunakan pada penelitian mencakup seluruh tahapan dari pengumpulan data, pengolahan, hingga evaluasi pada hasil yang diperoleh. Secara terstruktur, alur analisis data yang dilakukan tercakup dalam Gambar 1 berikut ini.



Gambar 1. *Flowchart* Penelitian

Penelitian ini menggunakan bahasa pemrograman *Python* dengan pustaka utama meliputi *Sastrawi*, *indoNLP*, dan *nltk* untuk proses tokenisasi, *stemming*, *lemmatization*, serta penghapusan *stopwords*. Representasi teks dilakukan menggunakan *TfidfVectorizer* dari *Scikit-learn* dan *word embedding GloVe* dari *Gensim* API. Algoritma LDA diimplementasikan menggunakan *Gensim*, sedangkan *K-Means* dan evaluasi *coherence score* menggunakan *Scikit-learn* dan *Gensim*. Tahapan ini akan dijelaskan secara rinci di bawah ini.

A. Pengumpulan Data

Dataset penelitian ini dikumpulkan melalui proses *scraping* komentar video YouTube yang dilakukan menggunakan bahasa pemrograman Python. Video tersebut secara khusus menghadirkan sosok publik figur bernama Arra. Proses *scraping* dilakukan dengan menggunakan API YouTube yang memungkinkan pengambilan data komentar secara lebih terstruktur dan sah, berbeda dengan metode web *scraping* yang biasa dilakukan dengan mengakses halaman HTML secara langsung. Penggunaan API mempermudah pengumpulan data tanpa melanggar kebijakan YouTube. Data yang diperoleh disimpan dalam bentuk format CSV dengan kolom utama berisi isi komentar. Data yang digunakan terdiri dari beberapa sumber data, dengan total dokumen yang berhasil terambil sebanyak 6.817.

TABEL 1
SUMBER DATA *SCRAPING* KOMENTAR YOUTUBE

No.	URL Video	*Channel*	Jumlah Komentar
1.	https://www.YouTube.com/watch?v=94x-LdIIjp4	TRANS TV Official – Pagi Pagi Ambyar	1.158
2.	https://www.YouTube.com/watch?v=JPMmyhn6vhM	TRANS TV Official – Brownis	1.503
3.	https://www.YouTube.com/watch?v=NFnm6sZkxfI	CURHAT BANG Denni Sumargo	3.141
4.	https://www.YouTube.com/watch?v=TIHECYJ9c4s	Feni Rose Official – Orang Tua Arra Angkat Bicara	1.393
5.	https://www.YouTube.com/watch?v=Kk9uI_usiKE	dr. Richard Lee, MARS	614
6.	https://www.YouTube.com/watch?v=kqaEu4tmmWY	TRANS7 OFFICIAL	536
7.	https://www.YouTube.com/watch?v=_YS2Y__cuDc	insertlive	122
8.	ttps://www.YouTube.com/watch?v=5sFOBtpvVAw	Feni Rose Official – Ara Bocah Kritis	507

B. Preprocessing

Preprocessing data teks bertujuan untuk mengurangi kebisingan dan membersihkan data agar dapat digunakan untuk analisis lanjutan [11]. Tahapan ini penting mengingat komentar YouTube sering kali mengandung kata tidak baku, singkatan, serta simbol-simbol yang tidak relevan. Pada penelitian ini, penanganan kata tidak baku mencakup berbagai bentuk slang dan penulisan fonetik yang umum digunakan dalam percakapan informal di media sosial. Adapun total token dalam korpus setelah *preprocessing* yaitu 121.394. Langkah-langkah *preprocessing* yang dilakukan dalam penelitian ini adalah sebagai berikut:

1) Penyaringan Kualitas Komentar

Langkah pertama yang dilakukan yaitu menjamin kualitas data komentar yang akan diproses. Komentar kosong dan komentar duplikat akan dihapus agar menghindari potensi bias dalam data. Pada penelitian ini tidak ada penghapusan komentar *toxic* karena video Arra sendiri banyak mengandung kata-kata kasar dan hal tersebut yang turut menjadi fokus penelitian ini.

2) Normalisasi Teks dan Slang

Langkah awal adalah memperbaiki kata-kata yang ditulis secara tidak baku atau bentuk slang [12]. Proses ini mencakup penggantian beberapa kata seperti “yg”, “y g”, dan “Yg” menjadi “yang”, serta koreksi terhadap penulisan yang berulang seperti “yaaaa” menjadi “ya”. Kata-kata lain seperti “woii” diubah menjadi “woy”, dan kata dengan kesalahan pengetikan seperti “adap” diperbaiki menjadi “adab”. Normalisasi ini membantu menyamakan bentuk kata yang berbeda namun memiliki makna yang sama.

3) Konversi ke Huruf Kecil (Case Folding)

Semua karakter dalam komentar diubah ke huruf kecil untuk menyamakan representasi kata yang sama namun ditulis dalam kapitalisasi berbeda, seperti “Adab” dan “adab” [13].

4) Pembersihan Simbol, Emoji, dan Angka

Karakter-karakter khusus seperti tanda seru, tanda tanya, emoji, dan simbol lainnya dihapus dari teks [14]. Pada proses ini, tanda koma tetap dipertahankan untuk mempertahankan struktur frasa tertentu. Selain itu, angka juga dihapus karena dinilai tidak memberikan makna signifikan dalam konteks analisis opini.

5) Penghapusan Stopwords

Tahapan ini dilakukan untuk menghapus kata-kata yang tidak memberikan informasi penting, seperti kata sambung atau kata ganti [13]. Penghapusan *stopwords* dilakukan dengan menggunakan daftar *stopwords* Bahasa Indonesia dan menambahkan secara manual kata “ya” dan “nya” yang sering muncul namun tidak mempengaruhi isi opini.

6) Lematisasi

Tahapan ini dilakukan dengan mengembalikan teks yang sudah bersih ke bentuk dasarnya (*lemma*) menggunakan lematisasi Bahasa Indonesia dengan bantuan *indoNLP* [14]. Misalnya kata “mengurus” dan “pengurus” menjadi urus.

7) Tokenisasi

Tokenisasi merupakan proses memecah teks menjadi unit-unit yang lebih kecil yang dikenal sebagai token. Token bisa terdiri dari kata, frasa, atau simbol yang memiliki arti tertentu dalam teks [13]. Pada tahap tokenisasi, teks yang sudah melalui proses normalisasi, konversi huruf kecil, pembersihan simbol, dan angka, akan dibagi menjadi kata-kata terpisah.

TABEL 2
HASIL NORMALISASI TEKS

Input	*Output*
Masih kecil cara bicaranya mulutnya kaya orang dewasa 😂	bicara mulut orang dewasa
Arra anak pintar org tua nya keren bisa mendidik anak	ara anak pintar orang tua bagus didik anak
Es krim gk pke es😂	es krim tidak pakai es
Seru😊 😅 😂	seru
plis jangan diundang lagi	tolong undang
lucuuu arra	lucu ara
Seru banget ngobrol sm ara 😅	seru obrol ara

Masyaa Allah... Tabarakallah...\nSeneng bangeet liat percakapan Mbak Feni Rose dg Arra yg smart, cantik dan lucuuu 😂😂❤❤ jd anak yg shalihah ya Arra sayang❤❤	senang lihat cakap rose ara pintar cantik lucu anak shalihah ara sayang
@@fadilahsvlog1818 budayakan nonton sampe habis didengerin noh pake telinganya, habis didengerin sampe habis baru deh lo komen. Kalo lu nonton sampe habis, kaga bakalan lu komen julid nyinyir, yg ada elo bakalan kagum sm kecerdasan ara. Keep smile and be positif thinking well...	budaya tonton habis dengar noh pakai telinga habis dengar habis komen tonton habis kagak komen julid nyinyir kagum pintar ara keep smile be positif thinking well
Gak ada lucu nya ngeselin iya	tidak lucu kesal

C. Text Representation

Representasi teks merupakan tahap penting dalam pemrosesan bahasa alami karena mengubah data teks mentah menjadi bentuk numerik yang dapat diproses oleh algoritma *machine learning.* Pada penelitian ini digunakan dua pendekatan utama dalam representasi teks yaitu *Term Frequency-Inverse Document Frequently* (TF-IDF) dan *Global Vectors for Word Representation (GloVe).*

1) Bag of Words (BoW)

BoW merupakan salah satu metode paling dasar dalam representasi teks. Model ini merepresentasikan dokumen sebagai kumpulan kata tanpa memperhatikan urutan atau konteksnya, dengan menghitung frekuensi kemunculan kata-kata dalam dokumen [15]. Sayangnya, BoW tidak mampu membedakan kata-kata yang sering muncul di banyak dokumen dan kata-kata yang spesifik untuk dokumen tertentu. Hal ini menyebabkan representasi yang dihasilkan kurang informatif dan berpotensi memunculkan *noise* dalam analisis data.

2) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF adalah metode berbasis statistik dan termasuk bagian dari teknik leksikal yang menghitung bobot penting sebuah kata terhadap dokumen dalam sebuah korpus. Bobot dihitung berdasarkan frekuensi kata dalam dokumen tertentu (*term frequency*) dan kebalikannya dari jumlah dokumen yang mengandung kata tersebut (*inverse document frequency*) [16]. Rumus yang digunakan sebagai berikut:

$$w_{ij} = tf_{ij} x \log\left(\frac{N}{df_j}\right) \quad (1)$$

dengan w_{ij} adalah bobot kata ke-i pada dokumen ke-j, tf_{ij} adalah frekuensi kemunculan kata, df_j adalah jumlah dokumen yang mengandung kata tersebut, dan N adalah total jumlah dokumen. Metode ini efektif untuk menyoroti kata-kata yang spesifik dalam suatu dokumen, namun bersifat *sparse* (jarang terisi) dan tidak menangkap konteks semantik [17]. Meskipun demikian, TF-IDF banyak digunakan dalam analisis teks klasik karena kesederhanaan dan efisiensinya.

3) Global Vectors for Word Representation (GloVe)

GloVe merupakan metode representasi teks berbasis *embedding* yang dilatih untuk menangkap hubungan semantik antar kata [18]. Berbeda dengan TF-IDF yang hanya mempertimbangkan frekuensi kata, *GloVe* membangun representasi vektor kata berdasarkan matriks *co-occurrence* global, yaitu seberapa sering kata-kata muncul bersama dalam konteks tertentu di seluruh korpus.

Model *GloVe* menghasilkan vektor berdimensi tetap (misalnya 100 atau 300 dimensi) yang menyimpan informasi semantik [19]. Kata-kata yang memiliki makna serupa akan memiliki vektor yang berdekatan secara geometris. Sebagai contoh, kata "adab" memiliki vektor yang dekat dengan "sopan", "etika", dan "moral", berdasarkan distribusi kemunculannya dalam korpus.

D. Topic Modeling

Topic Modeling adalah salah satu teknik *unsupervised machine learning* yang digunakan untuk menemukan topik-topik utama dari sekumpulan dokumen (*corpus*) [20]. Tujuan dari *topic modeling* adalah menemukan topik-topik tersembunyi di balik kumpulan kalimat atau dokumen. Pada penelitian ini digunakan dua metode utama dalam *Topic Modeling* yaitu, *Latent Dirichlet Allocation (LDA*) dan *K-Means*.

1) Latent Dirichlet Allocation (LDA)

LDA termasuk ke dalam keluarga model probabilistik, di mana setiap topik direpresentasikan sebagai distribusi probabilitas dari kemunculan kata-kata tertentu dalam korpus. Setiap dokumen dalam *LDA* dipandang sebagai kombinasi (campuran) dari berbagai topik, sehingga satu dokumen dapat mencermikan beberapa topik sekaligus dalam proporsi yang berbeda [20]. *LDA* dibangun atas asumsi *Bag of Words (*BoW*)*, di mana setiap dokumen dianggap sebagai kumpulan kata tanpa memperhatikan urutan kata dalam kalimat

[21]. Asumsi lain yang mendasari *LDA* adalah independensi antar topik, yang berarti bahwa topik-topik dianggap tidak saling bergantung [22]. Pada penelitian ini, akan dikaji LDA menggunakan model dasar dengan fitur hasil BoW dan TF-IDF untuk melihat seberapa efektif masing-masing Teknik dalam menangkap informasi yang relevan dalam teks. Disamping itu, hal ini berguna dalam membandingkan sejauh mana kedua metode ini dapat membantu dalam menemukan struktur topik yang tersembunyi.

2) K-Means

K-Means adalah algoritma *clustering* yang bekerja dengan memulai dari jumlah *cluster* (*k*) yang ditentukan secara bebas oleh pengguna. Setelah menentukan jumlah *cluster* yang digunakan, data yang telah direpresentasikan ke dalam bentuk vector akan dipartisi ke dalam *k* kelompok berdasarkan kedekatannya terhadap titik pusat *(centroid)* masing-masing *cluster* [20]. *Centroid* tersebut kemudian dihitung ulang secara iteratif untuk meminimalkan total jarak dalam satu *cluster* [23]. Proses ini memastikan bahwa anggota-anggota dalam satu *cluster* memiliki kemiripan yang tinggi satu sama lain. Pada penelitian ini, akan dikaji pendekatan representasi fitur menggunakan TF-IDF atau *GloVe*. Hal ini berguna dalam melihat seberapa berperan fitur yang direpresentasikan secara semantik dibandingkan fitur yang hanya memperhatikan frekuensi dan bobot kepentingan kata dalam dokumen.

E. Evaluasi Akhir

Pada tahapan ini, akan dilihat apakah topik-topik yang diperoleh model mempresentasikan isi dan struktur dari dokumen yang dianalisis. Model yang akan dibandingkan yaitu LDA model dasar, LDA dengan TF-IDF, *K-Means* dengan TF-IDF, dan *K-Means* dengan *GloVe*. Pada penelitian ini, metrik evaluasi yang akan digunakan yaitu *Coherence score* dengan rumus C_v. Pada penelitian ini, jumlah topik yang diuji divariasikan pada rentang 2 hingga 39 untuk mengevaluasi nilai C_v terbaik. Pemilihan rentang tersebut mengacu pada jumlah dokumen yang tersedia serta mengikuti praktik umum dalam pemodelan teks pendek untuk mengukur koherensi topik [24]. Pada pemodelan topik, koherensi topik mengukur kualitas data dengan membandingkan kesamaan semantik antara kata-kata yang berulang dalam suatu topik [25]. *Coherence score* adalah skala dari 0 hingga 1 dimana koherensi yang baik (kemiripan tinggi) memiliki skor 1 dan koherensi yang buruk (kemiripan rendah) memiliki skor 0 [26].

Rumus C_v mengukur koherensi topik berdasarkan kesamaan antara kata-kata dalam topik yang sama, dengan mempertimbangkan kata-kata yang sering muncul dalam konteks yang sama [27]. Formula untuk menghitung C_v sebagai berikut.

$$C_v = \frac{(\sum_{k=1}^{K} \sum_{n=1}^{N} \varphi S_i\ (\vec{u}, \vec{w}))}{Nx\ K} \tag{2}$$

Keterangan:

- $\varphi S_i\ (\vec{u}, \vec{w})$: *cosine similarity* antara dua vektor
- K: jumlah total topik pada dataset
- k: indeks untuk mengidentifikasi topik tertentu pada korpus
- N: jumlah total kata pada setiap topik
- n: indek untuk mengidentifikasi kata tertentu pada topik

Adapun *cosine similarity* dihitung dengan formula sebagai berikut.

$$\varphi S_i\ (\vec{u}, \vec{w}) = \frac{u_i . w_j}{\|u_i\| \|w_i\|} \tag{3}$$

Keterangan:

- $u_i . w_j$: perkalian dot antara vektor u dan vektor w
- $\|u_i\|$ dan $\|w_i\|$: panjang (norma) dari vektor u dan vektor w

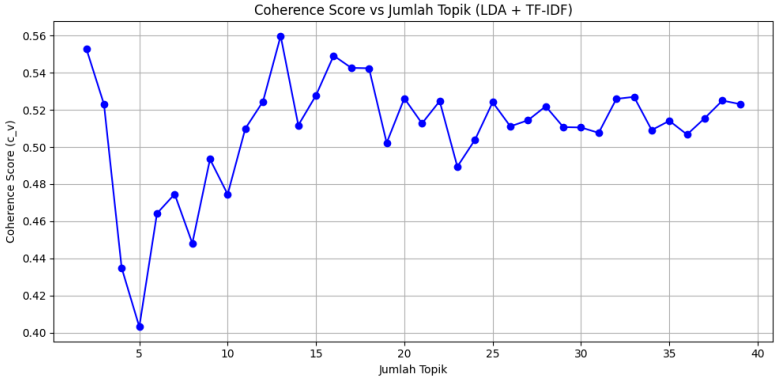
Perhitungan C_v dilakukan dengan beberapa langkah. Pertama, melakukan pemilihan kata-kata teratas dalam sebuah topik dari model LDA, dalam hal ini 10 kata yang paling berkontribusi. Selanjutnya, hitung kesamaan semantik antara setiap pasangan kata dengan menggunakan *cosine similarity* berdasarkan *word embeddings* kata. Setelah itu, koherensi topik dihitung dengan menjumlahkan skor kesamaan semantik dari semua pasangan kata dan menghitung rata-ratanya. Hasil akhirnya adalah C_v, yang bernilai antara 0 hingga 1.

III. HASIL DAN PEMBAHASAN

Pada penelitian ini, data dilakukan *preprocessing* yang khusus untuk data teks secara komprehensif. Setelah data berhasil dibersihkan, akan dilakukan lima pemodelan untuk mendapatkan nilai *coherence score* terbaik mengenai kasus pemodelan topik pada komentar YouTube Arra. Masing-masing model akan dikaji lebih dalam pada penjelasan berikut.

A. Pemodelan Topik dengan LDA Menggunakan BoW (LDA Dasar)

Model LDA berbasis *Bag of Words* (BoW) dibangun dengan menyesuaikan parameter untuk menghasilkan model topik yang optimal. Model terbaik diperoleh melalui proses *parameter tuning* pada dua parameter utama yaitu *alpha* dan *eta*. Parameter *alpha* mengontrol distribusi topik dalam setiap dokumen, nilai alpha yang rendah menghasilkan distribusi yang tersebar ke dalam sedikit topik, sedangkan nilai tinggi memungkinkan dokumen mencakup lebih banyak topik. Sementara itu, *eta* mengatur distribusi kata dalam setiap topik; nilai rendah menyebabkan topik hanya berisi sedikit kata dominan, sedangkan nilai tinggi membuat distribusi kata lebih merata dalam topik. Pada penelitian ini, kedua parameter diatur secara otomatis menggunakan opsi 'auto' pada paket Gensim, agar penyesuaian dilakukan berdasarkan karakteristik data. Selain tuning yang dilakukan secara otomatis, model ini melakukan evaluasi performa model pada jumlah topik yang berbeda yakni dalam rentang topik 2 hingga 39. Pada setiap jumlah topik, model LDA dilatih dengan 15 iterasi (*passes*) agar konvergensi stabil. Evaluasi model dilakukan menggunakan metrik *coherence score* (C_v), yang menilai seberapa koheren kata-kata dalam suatu topik.



Gambar 2. Hasil *Coherence score* pada Masing-masing Jumlah Klaster

Berdasarkan hasil analisis *Latent Dirichlet Allocation* (LDA) dengan *Bag of Words* (BoW), model berhasil mengidentifikasi 13 topik utama dalam dataset dengan *coherence score* terbaik sebesar 0.5598, yang menunjukkan bahwa topik-topik yang ditemukan cukup koheren dan relevan dengan teks yang dianalisis. Skor *coherence score* ini dipengaruhi oleh distribusi kata-kata dalam topik yang terbentuk dan seberapa relevan kata-kata tersebut muncul bersamaan dalam konteks yang serupa. Setiap topik memiliki distribusi kata yang cukup jelas dan mendefinisikan tema-tema yang spesifik.

TABEL 3
HASIL *TOPIC MODELING* MENGGUNAKAN LDA MODEL DASAR

Nomor Topik	Korpus
Topik 1	0.268*"ara" + 0.136*"pintar" + 0.042*"lucu" + 0.027*"anak" + 0.026*"cantik" + 0.024*"semoga" + 0.020*"semangat" + 0.017*"lihat" + 0.017*"sayang" + 0.016*"sehat"
Topik 2	0.024*"pendek" + 0.020*"lagu" + 0.018*"nyanyi" + 0.014*"raffi" + 0.014*"acara" + 0.014*"have" + 0.013*"bilang" + 0.013*"minus" + 0.012*"talk" + 0.012*"kata"
Topik 3	0.085*"main" + 0.030*"tuju" + 0.022*"bunda" + 0.020*"emak" + 0.017*"nikah" + 0.016*"doa" + 0.014*"jalan" + 0.014*"pasang" + 0.014*"rezeki" + 0.012*"pinternya"
Topik 4	0.212*"ibu" + 0.181*"ayah" + 0.029*"ilmu" + 0.024*"adab" + 0.012*"ara" + 0.012*"benar" + 0.009*"kagum" + 0.009*"adek" + 0.008*"terimakasih" + 0.007*"perhati"
Topik 5	0.069*"tonton" + 0.037*"tangis" + 0.020*"terbaik" + 0.020*"me" + 0.019*"podcast" + 0.018*"ayu" + 0.017*"pick" + 0.017*"cita" + 0.014*"racun" + 0.012*"emaknya"
Topik 6	0.065*"lala" + 0.040*"bocil" + 0.026*"salam" + 0.017*"tangan" + 0.016*"natural" + 0.015*"cium" + 0.014*"membully" + 0.013*"tiktok" + 0.012*"kerudung" + 0.011*"oh"
Topik 7	0.105*"anak" + 0.080*"orang" + 0.061*"tidak" + 0.042*"tua" + 0.019*"ajar" + 0.013*"didik" + 0.013*"omong" + 0.013*"dewasa" + 0.011*"salah" + 0.011*"tau"
Topik 8	0.074*"undang" + 0.036*"seru" + 0.026*"tv" + 0.014*"pickme" + 0.012*"ramah" + 0.011*"fyp" + 0.010*"penting" + 0.008*"wawancara" + 0.008*"bareng" + 0.007*"tolong"

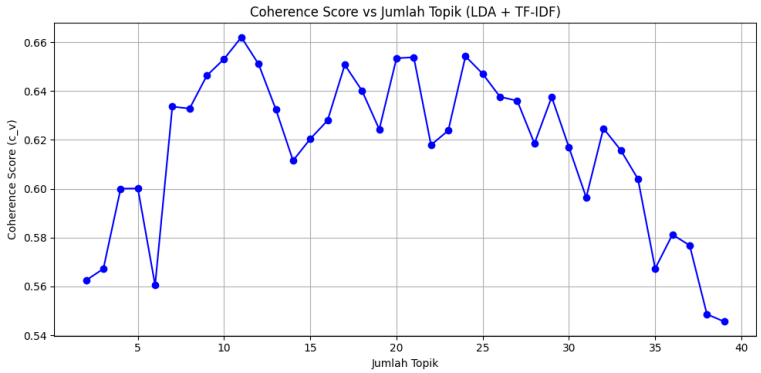
Topik 9	0.031*"rara" + 0.026*"bahagia" + 0.022*"tuhan" + 0.020*"mata" + 0.012*"mulu" + 0.011*"ros" + 0.010*"podcast" + 0.009*"richard" + 0.009*"si" + 0.009*"rambut"
Topik 10	0.029*"cari" + 0.024*"tidak" + 0.021*"kerja" + 0.021*"om" + 0.016*"ra" + 0.014*"pakai" + 0.013*"sosmed" + 0.012*"dokter" + 0.012*"surga" + 0.012*"jilbab"
Topik 11	0.053*"tasya" + 0.023*"songong" + 0.021*"angie" + 0.019*"duluan" + 0.018*"akhirat" + 0.015*"lebay" + 0.014*"duluan" + 0.014*"sedangkan" + 0.014*"harga" + 0.011*"abe"
Topik 12	0.018*"kocak" + 0.018*"benci" + 0.017*"super" + 0.014*"dikit" + 0.014*"jahat" + 0.013*"hingga" + 0.013*"pagi" + 0.013*"tahu" + 0.013*"tayang" + 0.010*"resiko"
Topik 13	0.025*"aunty" + 0.013*"sunda" + 0.011*"be" + 0.009*"menang" + 0.009*"would" + 0.008*"kids" + 0.008*"are" + 0.008*"nanti" + 0.008*"ekspresi" + 0.008*"life"

Analisis topik mengungkap beberapa tema dominan dalam korpus. Topik 1 yang didominasi kata "lucu", "cantik", dan "sayang" merepresentasikan ekspresi emosi positif dalam interaksi sosial. Topik 3 yang mencakup kata-kata seperti "main", "tujuh", "emak", dan "nikah" lebih berkaitan dengan tema keluarga dan kegiatan sosial**.** Topik 2 dan 5 dengan kata kunci "lagu", "nyanyi", dan "podcast" mengidentifikasi konten hiburan musik, sedangkan Topik 4 dan 7 yang mengandung "anak", "orang tua", dan "didik" jelas merefleksikan diskusi seputar parenting dan pendidikan. Pola menarik juga terlihat pada Topik 8 dan 12 dengan kata “pickme”, "seru", "kocak", dan "fyp" yang menjadi penanda konten media sosial, gaya hidup remaja, dan kebencian terkait Arra, serta Topik 13 yang memadukan kata bahasa Inggris ("kids", "life") dengan istilah lokal ("sunda") menunjukkan karakteristik konten bilingual dan pembahasan daerah Arra. Keterbatasan utama terletak pada munculnya beberapa kata yang kurang jelas konteksnya, yang mungkin disebabkan oleh karakteristik bahasa informal dalam korpus atau keterbatasan pendekatan BOW dalam menangkap kepentingan kata.

B. Pemodelan Topik dengan LDA Menggunakan TF-IDF

Model LDA tidak hanya dibangun menggunakan representasi *Bag-of-Words* (BoW), tetapi penelitian ini mencoba memperluas menggunakan TF-IDF sebagai bobot kata. Pendekatan ini penting untuk menurunkan pengaruh kata-kata umum yang sering muncul di banyak dokumen, namun tidak memberikan informasi spesifik terhadap topik tertentu. Langkah awal dilakukan dengan memecah komentar yang telah dipra-proses menjadi token, yaitu daftar kata yang telah dibersihkan dari slang, huruf berulang, simbol, dan *stopwords*. Kamus dan representasi BoW kemudian dibentuk dari token tersebut lalu ditransformasikan menjadi TF-IDF untuk memberikan bobot lebih pada kata-kata yang penting dalam konteks dokumen.

Proses pelatihan model LDA dilakukan dengan mengevaluasi berbagai jumlah topik, dimulai dari 2 hingga 39 topik. Untuk setiap jumlah topik, dilakukan pelatihan model menggunakan parameter α (alpha) dan η (eta) yang diatur secara otomatis melalui opsi 'auto', yang memungkinkan penyesuaian distribusi topik sesuai karakteristik data. Setiap model yang dihasilkan kemudian dievaluasi menggunakan *Coherence score* dengan metrik c_v.



Gambar 3. Hasil *Coherence score* pada Masing-masing Jumlah Klaster

Hasil pemodelan menunjukkan bahwa LDA dengan TF-IDF menghasilkan kualitas topik yang lebih baik dibanding LDA dengan BoW, hal ini terlihat dari peningkatan *coherence score* sebesar 18.2% menjadi 0.6620. Peningkatan ini terjadi karena TF-IDF secara efektif mampu mengurangi dominasi kata-kata umum yang kurang informatif melalui mekanisme *inverse document frequency*, serta memperkuat representasi kata-kata kunci spesifik melalui pembobotan *term frequency*. Hasilnya, setiap topik menjadi lebih terdefinisi dengan jelas, seperti terlihat pada Topik 3 (motivasi dan psikologi positif) yang didominasi kata "semangat", "bijak", dan "bahagia", serta Topik 6 (parenting) dengan fokus pada "anak", "orang tua", dan "pintar". Keunggulan utama TF-IDF terletak pada kemampuannya menghasilkan topik yang lebih interpretable. Dibandingkan dengan BOW yang

masih memasukkan kata-kata kurang relevan seperti "tidak" dan "tau" dalam Topik 6, TF-IDF berhasil mempertahankan fokus pada kata kunci yang benar-benar representatif.

TABEL 4
HASIL *TOPIC MODELING* MENGGUNAKAN LDA DENGAN TF IDF

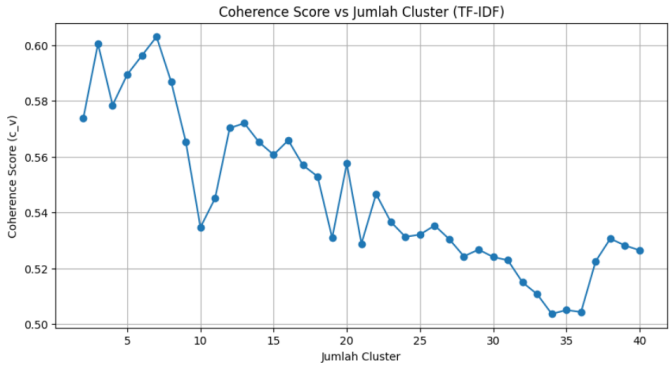
Nomor Topik	Korpus
Topik 1	0.015*"me" + 0.015*"pick" + 0.009*"hijab" + 0.006*"thinking" + 0.006*"judul" + 0.005*"abang" + 0.005*"ekonomi" + 0.004*"lahir" + 0.004*"ponakan" + 0.004*"logat"
Topik 2	0.006*"konsep" + 0.006*"suami" + 0.006*"perhati" + 0.006*"pasang" + 0.005*"YouTube" + 0.005*"an" + 0.005*"tahu" + 0.005*"tabarakalloh" + 0.005*"harga" + 0.004*"pandang"
Topik 3	0.020*"seru" + 0.012*"pinternya" + 0.009*"akhirat" + 0.009*"sebal" + 0.008*"menit" + 0.005*"bahaya" + 0.005*"publik" + 0.004*"negara" + 0.004*"ain" + 0.003*"empati"
Topik 4	0.040*"semangat" + 0.010*"bijak" + 0.009*"salim" + 0.008*"bahagia" + 0.007*"kuat" + 0.007*"buzzer" + 0.007*"cium" + 0.006*"ulang" + 0.005*"pagi" + 0.005*"duduk"
Topik 5	0.010*"komentar" + 0.009*"bocil" + 0.009*"indonesia" + 0.009*"arra" + 0.007*"rose" + 0.007*"kritis" + 0.007*"kritik" + 0.006*"bangga" + 0.006*"sifat" + 0.006*"ayu"
Topik 6	0.008*"celoteh" + 0.007*"korea" + 0.007*"pickme" + 0.007*"bela" + 0.006*"sholeh" + 0.006*"duit" + 0.006*"jawab" + 0.006*"kagum" + 0.006*"tipe" + 0.005*"bentar"
Topik 7	0.048*"ara" + 0.037*"pintar" + 0.029*"anak" + 0.023*"orang" + 0.019*"tidak" + 0.019*"lucu" + 0.017*"tua" + 0.012*"ibu" + 0.011*"cantik" + 0.011*"bagus"
Topik 8	0.011*"teteh" + 0.009*"bunda" + 0.008*"pabrik" + 0.008*"kesini" + 0.007*"komennya" + 0.006*"bubar" + 0.005*"gemar" + 0.005*"contoh" + 0.005*"kulit" + 0.005*"tukang"
Topik 9	0.034*"sehat" + 0.021*"bocah" + 0.009*"jual" + 0.007*"lebay" + 0.007*"asli" + 0.006*"susah" + 0.006*"surga" + 0.006*"bohong" + 0.005*"cucu" + 0.005*"semangat"
Topik 10	0.012*"ra" + 0.009*"raffi" + 0.008*"bosan" + 0.008*"mantap" + 0.007*"sunda" + 0.007*"muncul" + 0.005*"kemarin" + 0.005*"cermin" + 0.005*"juta" + 0.005*"jehan"
Topik 11	0.010*"gemes" + 0.006*"wajah" + 0.006*"zehan" + 0.005*"jaman" + 0.004*"uji" + 0.004*"dengki" + 0.004*"muji" + 0.004*"buat" + 0.004*"maksud" + 0.004*"dosa"

Topic modeling menggunakan LDA dengan pendekatan TF-IDF menghasilkan beberapa tema prioritas pada masing-masing sebaran topik. Topik 1 (Identitas & Interaksi Sosial), topik ini didominasi kata "me", "pick", dan "hijab" menunjukkan pembahasan tentang identitas personal (khususnya perempuan muslim). Topik 2 (Kehidupan Rumah Tangga & Konten Digital), didasarkan pada kata "suami", "perhati", dan "pasang" merefleksikan dinamika rumah tangga, sementara "YouTube" dan "harga" menunjukkan keterkaitan dengan konsumsi konten digital dan ekonomi domestik. Topik 3 (Konten Hiburan & Isu Sosial) didominasi oleh kata "seru" dan "pinternya" menjadi penanda konten hiburan, sedangkan "akhirat", "negara", dan "empati" menunjukkan nilai-nilai religius dan sosial. Topik 4 (Motivasi) didominasi oleh kata "semangat" (0.040) dengan dukungan kata "bijak" dan "bahagia" membentuk narasi pengembangan diri. Pada topik 5 (Diskusi Publik & Identitas Nasional) terdapat kombinasi "komentar", "kritis", dan "Indonesia" menunjukkan topik tentang tanggapan masyarakat terhadap isu nasional. Topik 6 yang membahas Budaya Pop & Nilai Religius didasarkan pada kata "korea", "pickme", dan "sholeh" menciptakan polaritas menarik antara budaya pop global dengan nilai lokal-religius. Topik 7 mengenai Parenting & Pendidikan Anak sebagai topik paling dominan (alpha tertinggi) yang merepresentasikan kata "ara", "pintar", "anak", dan "orang tua" membentuk klaster yang solid tentang pengasuhan anak. Disamping itu, juga terdapat Topik 8 yang membahas Kehidupan Sehari-hari & Industri, Topik 9 mengenai Kesehatan & Gaya Hidup, Topik 10 mengenai Konten Lokal & Hiburan, serta Topik 11 mengenai Ekspresi Emosi & Moralitas.

LDA dengan TF-IDF terbukti lebih unggul daripada BOW dalam hal koherensi topik, spesifisitas kata kunci, dan interpretabilitas. Pemodelan LDA dengan TF-IDF mengungguli penelitian pemodelan topik pariwisata dibandingkan penelitian [9] yang menghasilkan nilai koherensi 0,613 dan 0,528.

C. Pemodelan Topik dengan K-Means Menggunakan TF-IDF

Pemodelan topik juga dilakukan menggunakan *K-Means* dengan fitur yang telah direpresentasikan melalui TF-IDF. Proses dimulai dengan tokenisasi dokumen dan pembuatan dictionary serta corpus BoW. Kemudian, model TF-IDF diterapkan untuk mengonversi corpus BoW menjadi representasi TF-IDF. Setelah itu, matriks numerik hasil TF-IDF dikonversi ke bentuk dense. Pada prosesnya, dilakukan tuning jumlah klaster terbaik mulai dari 2 hingga 39. Pada setiap jumlah *cluster*, model menghitung *coherence score* yang mengukur seberapa koheren kata-kata dalam setiap *cluster*. Skor *coherence* yang tinggi mengindikasikan jumlah klaster dalam pemodelan topik yang paling baik. Berikut ini plot hasil *coherence score* pada masing-masing jumlah klaster.



Gambar 4. Hasil *Coherence score* pada Masing-masing Jumlah Klaster

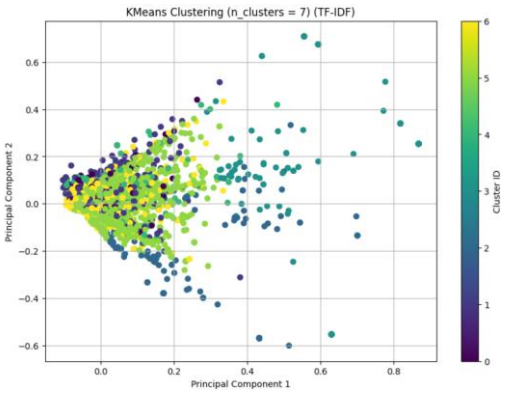
Hasil pemodelan topik menggunakan *K-Means* menunjukan 7 klaster dengan model terbaik, hal ini didasarkan pada *coherence score* tertinggi, mencapai 0,6029. Model *K-Means* berhasil menghasilkan topik yang koheren dan saling terkait dengan baik, meskipun nilainya sedikit lebih rendah dibandingkan dengan model LDA menggunakan TF-IDF (0.6620). Hasil ini menunjukkan bahwa topik yang dihasilkan kemungkinan lebih umum atau lebih luas, dibandingkan dengan hasil yang lebih terfokus pada LDA.

Salah satu perbedaan utama antara LDA dan *K-Means* adalah bahwa LDA adalah model probabilistik yang mencoba mengidentifikasi topik berdasarkan distribusi kata dalam dokumen, sedangkan *K-Means* menggunakan teknik pembelajaran berbasis jarak untuk mengelompokkan data ke dalam klaster tanpa asumsi distribusi kata-kata dalam topik. Setiap *cluster* yang dihasilkan dapat dianggap sebagai topik yang berisi kata-kata kunci yang memiliki kemiripan atau keterkaitan yang tinggi.

TABEL 5
HASIL PEMODELAN TOPIK MENGGUNAKAN *K-MEANS* DENGAN TF-IDF

Nomor Topik	Korpus
Topik 1	['ara', 'pintar', 'umur', 'tidak', 'pikir', 'omong', 'bicara', 'orang', 'anak', 'dewasa']
Topik 2	['parenting', 'bagus', 'salah', 'tidak', 'pintar', 'ajar', 'ara', 'anak', 'orang', 'tua']
Topik 3	['cerdas', 'suka', 'sayang', 'tidak', 'gemas', 'anak', 'cantik', 'pintar', 'ara', 'lucu']
Topik 4	['ibu', 'bagus', 'semoga', 'gemas', 'salihah', 'sehat', 'cantik', 'anak', 'pintar', 'ara']
Topik 5	['pintar', 'salah', 'bagus', 'ayah', 'ara', 'ibu', 'tua', 'orang', 'anak', 'didik']
Topik 6	['orang', 'gemas', 'semoga', 'cantik', 'bagus', 'ayah', 'pintar', 'ibu', 'anak', 'ara']
Topik 7	['omong', 'pintar', 'ilmu', 'ara', 'orang', 'ajar', 'anak', 'suka', 'adab', 'tidak']

Berdasarkan hasil analisis topik yang dihasilkan, sebagian besar topik berfokus pada tema pendidikan anak dan perasaan orang tua terhadap anak. Topik 1, dengan kata-kata seperti "pintar", "anak", dan "dewasa", membahas perbedaan perkembangan antara anak-anak dan orang dewasa, sementara topik 2 berfokus pada parenting dan cara orang tua mengasuh anak dengan nilai yang baik. Topik 3, dengan kata-kata seperti "sayang", "gemas", dan "lucu", menunjukkan perasaan kasih sayang terhadap anak-anak, sedangkan topik 4 lebih banyak berbicara tentang harapan orang tua untuk kesehatan dan perilaku positif anak. Topik 5 berkaitan dengan pendidikan anak dan peran orang tua dalam mendidik anak dengan cara yang baik. Terakhir, topik 6 berfokus pada pendidikan karakter dan adab anak, termasuk pengajaran nilai moral dan komunikasi yang baik. Secara keseluruhan, topik-topik ini mencerminkan perhatian besar terhadap perkembangan anak, pengasuhan yang baik, dan harapan positif orang tua terhadap anak. Berikut visualisasi dalam dua dimensi terkait *clustering K-Means* menggunakan TF-IDF pada pemodelan topik.



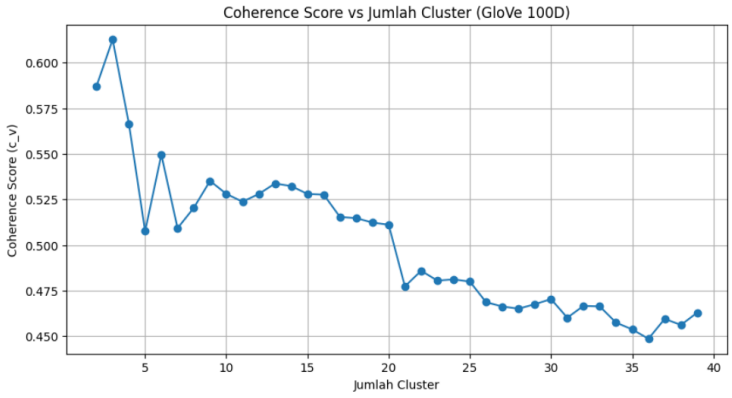
Gambar 5. Visualisasi 2D terkait *clustering K-Means* menggunakan TF-IDF

Visualisasi hasil *K-Means clustering* dengan tujuh *cluster*, yang direpresentasikan dalam dua dimensi melalui Principal Component Analysis (PCA) dari data TF-IDF, memperlihatkan pengelompokan data berdasarkan kemiripan fitur tekstualnya. Sebaran titik-titik berwarna pada scatter plot menunjukkan bahwa algoritma *K-Means* berhasil mengidentifikasi kelompok-kelompok yang berbeda, meskipun terdapat beberapa area tumpang tindih yang mengindikasikan kemiripan antar *cluster*. Informasi mengenai kata-kata teratas dalam setiap *cluster* memberikan wawasan tematik yang signifikan. Dengan demikian, visualisasi ini secara keseluruhan menggambarkan bagaimana dokumen atau teks yang direpresentasikan oleh TF-IDF berhasil dikelompokkan oleh *K-Means* berdasarkan kemiripan leksikal, dan informasi kata-kata teratas membantu kita memahami topik yang mendasari setiap kelompok tersebut.

D. Pemodelan Topik dengan K-Means Menggunakan GloVe

Penelitian ini berusaha melihat implementasi *GloVe* dengan 100D dan 300D sebagai teknik representasi teks yang mampu mengukur makna semantik dalam dokumen. Proses ini di awali dengan mengubah setiap dokumen menjadi vektor numerik berdasarkan rata-rata representasi kata dari model *GloVe* yang sudah dilatih (*pre-trained model*). Selanjutnya, dilakukan percobaan berbagai jumlah kelompok (*cluster*) untuk menemukan *cluster* terbaik dalam mengevaluasi kualitas pengelompokan tersebut menggunakan metrik koherensi topik (C_v). Jumlah *cluster* yang dicoba yaitu antara 2 sampai 39 *cluster*.

Berdasarkan *K-Means clustering* menggunakan *GloVe* 100D, model berhasil mengidentifikasi 3 *cluster* dengan *coherence score* terbaik sebesar 0.6127. Disamping itu, *K-Means clustering* menggunakan *GloVe* 300D berhasil menghasilkan 2 *cluster* terbaik dengan coherence scor hanya sebesar 0.5525. Ketika jumlah dimensi terlalu banyak, data yang lebih kompleks dapat mengarah pada overfitting atau membuat model kesulitan mengidentifikasi pola yang jelas. Penggunaan dimensi yang lebih rendah (seperti 100D) bisa lebih optimal untuk dataset yang lebih sederhana atau yang tidak memerlukan representasi kata yang sangat terperinci. *GloVe* 300D lebih cocok untuk dataset yang sangat besar dan kompleks, di mana lebih banyak dimensi dapat mengungkapkan hubungan yang lebih dalam antar kata. Pada kasus ini, peneliti menggunakan data komentar YouTube yang hanya berjumlah 6000-an dokumen serta tergolong dalam *short text*. Penelitian ini menyoroti bahwa dimensi yang lebih kecil mampu merepresentasikan makna kata dengan lebih baik, hal ini ditunjukan dengan *coherence score* yang lebih tinggi pada *GloVe* 100D.



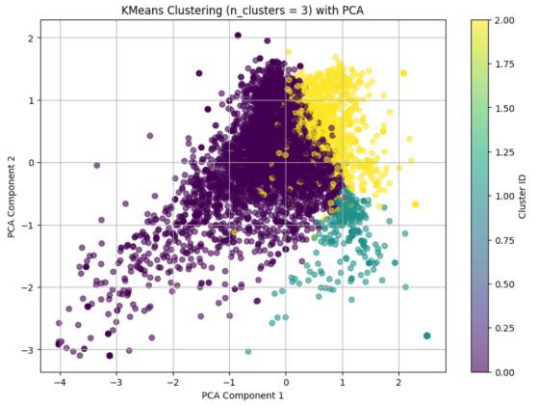
Gambar 6. Hasil *Coherence score* pada Masing-masing Jumlah Klaster

Coherence score yang tinggi pada model *K-Means* menggunakan *GloVe* 100D menunjukkan bahwa model berhasil membentuk *cluster* yang koheren, di mana kata-kata yang muncul dalam setiap *cluster* saling terkait dengan baik. Vektor *GloVe* mampu menangkap hubungan semantik antar kata, yang memungkinkan model untuk memahami kata-kata yang mirip secara semantik atau kata-kata yang sering muncul dalam konteks yang sama. Keunggulan *GloVe* terlihat dari kemampuannya menangani variasi kata-kata secara lebih efektif. Misalnya, kata-kata yang memiliki makna serupa meskipun ditulis dengan cara berbeda, seperti "belajar" dan "mempelajari", akan dianggap lebih mirip dalam ruang vektor *GloVe*, sedangkan TF-IDF tidak memperhitungkan hubungan ini secara eksplisit. Dengan demikian, *K-Means* dengan *GloVe* mampu menghasilkan *cluster* yang lebih koheren.

TABEL 6
HASIL PEMODELAN TOPIK MENGGUNAKAN *K-MEANS* DENGAN *GLOVE* 100D

Nomor Topik	Korpus
Topik 1	['orang', 'tidak', 'anak', 'ara', 'tua', 'pintar', 'ibu', 'ajar', 'ayah', 'suka']
Topik 2	['ara', 'pintar', 'lucu', 'orang', 'cantik', 'lihat', 'gemas', 'tidak', 'anak', 'tua']
Topik 3	['anak', 'orang', 'pintar', 'ara', 'tua', 'tidak', 'dewasa', 'ajar', 'didik', 'omong']

Hasil *clustering* menunjukkan tiga topik utama yang berkaitan dengan pendidikan anak dan peran orang tua dalam pengasuhan. Topik 1 berfokus pada pendidikan anak, dengan kata-kata seperti "anak", "ibu", "ayah", "ajar", dan "pintar" yang menekankan peran orang tua dalam mendidik anak-anak mereka, serta dinamika hubungan orang tua dengan anak, baik dalam konteks positif maupun negatif. Topik 2 lebih mengarah pada perasaan sayang dan kekaguman terhadap anak-anak, dengan kata-kata seperti "lucu", "gemas", "cantik", dan "anak" yang mencerminkan ekspresi kekaguman terhadap sifat menggemaskan atau perilaku anak-anak, serta pengamatan terhadap kecerdasan dan penampilan mereka. Topik 3 membahas perbandingan antara anak-anak dan orang dewasa, serta pentingnya pendidikan dalam perkembangan anak, dengan kata-kata seperti "dewasa", "didik", dan "ajar" yang menunjukkan peran orang tua atau pendidik dalam mendidik anak-anak agar berkembang menjadi pribadi yang pintar. Berikut visualisasi dalam dua dimensi terkait *clustering K-Means* menggunakan *GloVe* pada pemodelan topik.



Gambar 7. Visualisasi 2D terkait *clustering K-Means* menggunakan *GloVe* 100D

Terlihat adanya pemisahan yang jelas antara tiga *cluster* yang ditandai dengan warna yang berbeda (kuning, hijau, dan ungu). Pemisahan ini menunjukkan bahwa *K-Means* berhasil mengelompokkan dokumen-dokumen dalam tiga kelompok topik yang berbeda dengan baik. Ini berarti bahwa model berhasil menemukan pola yang terpisah dengan jelas di dalam data, dan *cluster-cluster* tersebut tidak tumpang tindih satu sama lain secara signifikan. Visualisasi ini mengindikasikan bahwa *K-Means clustering* dengan 3 *cluster* berhasil memisahkan dokumen-dokumen ke dalam tiga kelompok topik yang terpisah dengan jelas.

Berdasarkan kajian yang telah dipaparkan di atas, proses pemodelan *topic modeling* tidak bisa dilihat hanya dari kajian suatu formulasi maupun otomatisasi model melalui bahasa pemprograman. Manusia perlu mengkaji lebih dalam ketepatan pemodelan, apalagi pada kasus *unsupervised learning*. Berikut ini merupakan ringkasan keseluruhan performa model *topic modeling* dalam penelitian ini.

TABEL 7
RINGKASAN PERFORMA MODEL BERDASARKAN *COHERENCE SCORE*

No.	Model	*Coherence score*	Jumlah Topik
1.	LDA dengan BoW	0,5598	13
2.	LDA dengan TF-IDF	0,6620	11
3.	*K-Means* dengan TF-IDF	0,6029	7
4	*K-Means* dengan *GloVe* 100D	0,6127	3
5.	*K-Means* dengan *GloVe* 300D	0,5525	2

Pada Tabel 5 terlihat bahwa LDA dengan TF-IDF memiliki performa paling tinggi, yakni *coherence score* mencapai 0,6620, menunjukan model ini unggul dalam hal koherensi dibandingkan model lainnya. Meskipun demikian, penting untuk mempertimbangkan interpretabilitas dan kesesuaian pengelompokan topik secara praktis. Berdasarkan kajian yang dipaparkan, LDA dengan TF-IDF dengan 11 topik masih terjadi tumpang tindih yang signifikan pada masing-masing topik, sehingga interpretasi topik sulit dilakukan. Disamping itu, beberapa topik yang dihasilkan tidak relevan dengan pembahasan yang ada, yang mengindikasikan adanya kesalahan

pemaknaan pada topik-topik tersebut. Disisi lain, model *K-Means* dengan *GloVe* 100D menempati urutan kedua berdasarkan skor koherensi, dengan total topik sebanyak 3. Meskipun memiliki skor koherensi yang sedikit lebih rendah, model ini lebih unggul dalam hal interpretabilitas. Berdasarkan distribusi kata-kata per *cluster*, model ini sangat mudah untuk dipahami, didukung dengan visualisasi 2D yang menunjukan bahwa pemisahan antar *cluster* sangat jelas. Oleh sebab itu, pada kasus pemodelan topik pada komentar YouTube Arra, model *K-Means* dengan *GloVe* 100D menjadi pilihan yang tepat.

IV. SIMPULAN

Hasil kajian pada evaluasi efektivitas kombinasi model LDA dan *K-Means* dengan representasi teks leksikal (BoW dan TF-IDF) serta semantik (*GloVe*) untuk mengidentifikasi topik yang dominan dalam komentar YouTube Arra, menunjukkan bahwa LDA dengan TF-IDF menghasilkan nilai *coherence score* tertinggi sebesar 0.6620. Meskipun nilai koherensi yang diperoleh tinggi, topik-topik yang dihasilkan oleh model ini cenderung tumpang tindih antar makna dan kurang sesuai dengan konteks, sehingga menyulitkan proses interpretasi. Sebaliknya, *K-Means* dengan *GloVe* 100D yang menempati peringkat kedua dengan *coherence score* sebesar 0.6127, lebih unggul dalam hal interpretabilitas. Model ini membentuk 3 *cluster* yang terpisah dengan jelas. Keunggulan ini diperoleh karena kemampuan *GloVe* dalam mengenali hubungan semantik antar kata yang secara leksikal berbeda namun bermakna serupa. Dengan demikian, meskipun LDA dengan TF-IDF lebih tinggi nilai *coherence score*-nya, model *K-Means* dengan *GloVe* 100D memberikan hasil yang lebih baik untuk interpretasi topik. Keunggulan ini menjadikannnya metode yang baik untuk eksplorasi topik pada data sosial media yang kompleks, khususnya pada kasus pemodelan topik komentar YouTube Arra. Penelitian ini berada dalam ruang lingkup Teknik Informatika, khususnya pada cabang *text mining* dan pemrosesan bahasa alami (*natural language processing*). Hasil studi ini berkontribusi pada pengembangan metode analisis teks untuk data tidak terstruktur, serta dapat diterapkan secara praktis dalam sistem penyaringan isu sensitif, atau pengelompokan konten untuk sistem rekomendasi berbasis topik. Pengembangan penelitian ke depan dapat dilakukan dengan mengeksplorasi model berbasis *deep learning* seperti *BERTopic* atau *Top2Vec* yang mampu menangkap konteks semantik secara lebih baik agar tidak tumpang tindih. Validasi hasil model juga dapat ditingkatkan dengan melakukan *merging topic* untuk memperkuat interpretasi topik yang diperoleh dari model-model yang digunakan. Namun dalam penelitian ini, ingin disoroti performa masing-masing model secara *default* sehingga tanpa ada *merging topic* yang tumpang tindih. Penelitian juga dapat diperluas dengan membandingkan hasil pemodelan topik antar *platform* media sosial, sehingga diperoleh perbedaan karakteristik diskusi antar media sosial.

DAFTAR PUSTAKA

[1] kompasiana.com, “Mengenal Ara, Balita Viral di TikTok karena Hobi Deep Talk - Kompasiana.com.” Diakses: 10 April 2025. [Daring]. Tersedia pada: https://www.kompasiana.com/ahmadsidikx/669aa19ec925c4337655cc02/mengenal-ara-balita-viral-di-tiktok-karena-hobi-deep-talk

[2] N. A. Rakhmawati, R. B. Waskitho, D. A. Rahman, dan M. F. A. U. Nuha, “Klasterisasi Topik Konten Channel Youtube Gaming Indonesia Menggunakan Latent Dirichlet Allocation,” *JIEET*, vol. 5, no. 2, hlm. 78–83, Des 2021, doi: 10.26740/jieet.v5n2.p78-83.

[3] H. Jelodar *dkk.*, “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey,” 6 Desember 2018, *arXiv*: arXiv:1711.04305. doi: 10.48550/arXiv.1711.04305.

[4] G. Rosalinda, R. Santoso, dan P. Kartikasari, “PEMODELAN TOPIK ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN LATENT DIRICHLET ALLOCATION,” *J.Gauss*, vol. 11, no. 4, hlm. 554–561, Feb 2023, doi: 10.14710/j.gauss.11.4.554-561.

[5] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, dan J. Bian, “Evaluation of clustering and topic modeling methods over health-related tweets and emails,” *Artificial Intelligence in Medicine*, vol. 117, hlm. 102096, Jul 2021, doi: 10.1016/j.artmed.2021.102096.

[6] M. Das, S. Kamalanathan, dan P. J. A. Alphonse, “A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset,” *International Conference on Computational Linguistics and Intelligent Systems*, 2023, doi: 10.48550/arXiv.2308.04037.

[7] N. Badri, F. Kboubi, dan A. H. Chaibi, “Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection,” *Procedia Computer Science*, vol. 207, hlm. 769–778, 2022, doi: 10.1016/j.procs.2022.09.132.

[8] D. Andra dan A. B. Baizal, “E-commerce Recommender System Using PCA and K-Means Clustering,” *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 6, no. 1, hlm. 57–63, Feb 2022, doi: 10.29207/resti.v6i1.3782.

[9] N. L. P. M. Putu, Ahmad Zuli Amrullah, dan Ismarmiaty, “Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation,” *RESTI*, vol. 5, no. 1, hlm. 123–131, Feb 2021, doi: 10.29207/resti.v5i1.2587.

[10] R. Sabbagh dan F. Ameri, “A Framework Based on K-Means Clustering and Topic Modeling for Analyzing Unstructured Manufacturing Capability Data,” *Journal of Computing and Information Science in Engineering*, vol. 20, no. 1, hlm. 011005, Feb 2020, doi: 10.1115/1.4044506.

[11] C. Humam dan A. D. Laksito, “Implementasi Aplikasi Sentimen Pada Data Twitter Jelang Pemilu 2024,” *JPIT*, vol. 8, no. 2, hlm. 105–112, Mei 2023, doi: 10.30591/jpit.v8i2.5051.

[12] S. Khomsah dan A. S. Aribowo, “Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia,” vol. 4, no. 4, 2020.

[13] D. Rifaldi, Abdul Fadlil, dan Herman, “Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet ‘Mental Health,’” *Decode*, vol. 3, no. 2, hlm. 161–171, Apr 2023, doi: 10.51454/decode.v3i2.131.

[14] R. A. Naufal dan A. R. Pratama, “Analisis Sentimen terhadap Cyberbullying di Media Sosial dengan CrowdTangle,” *AUTOMATA*, 2023.

[15] S. Akuma, T. Lubem, dan I. T. Adom, "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets," *Int. j. inf. tecnol.*, vol. 14, no. 7, hlm. 3629–3635, Des 2022, doi: 10.1007/s41870-022-01096-4.
[16] Y. Fu dan Y. Yu, "Research on Text Representation Method Based on Improved TF-IDF," *J. Phys.: Conf. Ser.*, vol. 1486, no. 7, hlm. 072032, Apr 2020, doi: 10.1088/1742-6596/1486/7/072032.
[17] I. Nyoman Prayana Trisna, N. Wayan Emmy Rosiana Dewi, dan M. Alam Pasirulloh, "Oversampling vs. undersampling in TF-IDF variations for imbalanced Indonesian short texts classification," *TELKOMNIKA*, vol. 23, no. 2, hlm. 382, Apr 2025, doi: 10.12928/telkomnika.v23i2.26510.
[18] M. A. Ikfini M dan E. B. Setiawan, "Topic Detection on Twitter using GloVe with Convolutional Neural Network and Gated Recurrent Unit," *bits*, vol. 5, no. 2, Sep 2023, doi: 10.47065/bits.v5i2.4057.
[19] T. F. Abdillah, H. Hasmawati, dan B. Bunyamin, "Comparison of TF-IDF and GloVe Word Embedding for Sentiment Analysis of 2024 Presidential Candidates," *bits*, vol. 6, no. 2, hlm. 961–969, Sep 2024, doi: 10.47065/bits.v6i2.5668.
[20] B. Bengfort, R. Bilbro, dan T. Ojeda, "Applied Text Analysis with Python," *O'Reilly Media, Inc.*, 2018.
[21] D. M. Blei, A. Y. Ng, dan M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, 2003.
[22] D. Maier *dkk.*, "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology," *Communication Methods and Measures*, vol. 12, no. 2–3, hlm. 93–118, Apr 2018, doi: 10.1080/19312458.2018.1430754.
[23] M. Faisal, E. M. Zamzami, dan Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *J. Phys.: Conf. Ser.*, vol. 1566, no. 1, hlm. 012112, Jun 2020, doi: 10.1088/1742-6596/1566/1/012112.
[24] S. Syed dan M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," dalam *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, Japan: IEEE, Okt 2017, hlm. 165–174. doi: 10.1109/dsaa.2017.61.
[25] R. Albalawi, T. H. Yeap, dan M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front. Artif. Intell.*, vol. 3, hlm. 42, Jul 2020, doi: 10.3389/frai.2020.00042.
[26] D. Cline dan J. Ryan, "Exploring Coherence Metrics for Optimizing Topic Models of Humpback Song," 2020.
[27] Z. N. Muna, B. D. Setiawan, dan R. S. Perdana, "Penerapan Pemodelan Topik Komentar Melalui Media Sosial Twitter Menggunakan Latent Dirichlet Allocation (Studi Kasus: Pemerintah Kota Malang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2024.