JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

Analisis Topik Komentar Youtube pada Lagu Tema FIFA World Cup 2026 Menggunakan LDA

M. Dhafa Adjie Saputra[1], Fadhel Muhammad[2], Muhammad Rizky Pribadi[3]

[123]Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Palembang, Indonesia

Email: [1]mdhafaadjiesaputra_2327250081@mhs.mdp.ac.id, [2]fadhelmuhammad_2327250070@mhs.mdp.ac.id, [3]rizky@mdp.ac.id

Abstrak-Komentar yang ditinggalkan pengguna pada platform YouTube dapat dimanfaatkan untuk memahami berbagai respons publik terhadap suatu konten digital. Penelitian ini berfokus pada identifikasi pola pembahasan yang muncul pada komentar video musik Lighter yang digunakan sebagai lagu resmi FIFA World Cup 2026. Data penelitian berupa 398 komentar berbahasa Inggris diperoleh melalui proses *web scraping* menggunakan platform Apify. Sebelum dianalisis, data melalui serangkaian tahapan *preprocessing* yang mencakup pembersihan teks, tokenisasi, penghapusan *stopword*, pembentukan *bigram*, dan *lemmatization*. Proses ekstraksi topik dilakukan menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk menemukan kelompok pembahasan yang dominan dalam kumpulan komentar. Hasil pemodelan menunjukkan tiga tema utama yang berkaitan dengan penilaian terhadap kualitas musik, tanggapan mengenai kesesuaian lagu dengan atmosfer sepak bola, dan diskusi umum seputar video musik FIFA. Evaluasi menggunakan *coherence score* menghasilkan nilai 0,466 yang mengindikasikan bahwa topik yang terbentuk memiliki tingkat konsistensi yang cukup baik untuk diinterpretasikan. Temuan penelitian menunjukkan bahwa pendekatan LDA mampu digunakan sebagai metode yang efektif dalam mengidentifikasi kecenderungan pembahasan dan opini pengguna pada komentar YouTube berbasis teks pendek.

Kata Kunci: FIFA World Cup 2026; Komentar Youtube; Latent Dirichlet Allocation; Topic Modeling; Web Scraping

Abstract-*User-generated comments on YouTube provide valuable information for understanding public responses to digital content. This study investigates discussion patterns found in comments posted on the music video Lighter, the official theme song of the FIFA World Cup 2026. A total of 398 English-language comments were collected through a web scraping process using the Apify platform. Prior to analysis, the dataset underwent several preprocessing stages, including text cleaning, tokenization, stopword elimination, bigram generation, and lemmatization. Topic extraction was carried out using the Latent Dirichlet Allocation (LDA) method to uncover dominant themes within the comment collection. The results revealed three primary topics related to perceptions of music quality, opinions regarding the suitability of the song for a football event, and general discussions about FIFA music videos. Model evaluation produced a coherence score of 0.466, indicating a satisfactory level of semantic consistency among the generated topics. These findings demonstrate that LDA can effectively identify discussion trends and user opinions from short-text YouTube comments.*

Keywords*: FIFA World Cup 2026; Latent Dirichlet Allocation; Topic Modeling; Web Scraping; Youtube Comments*

1. PENDAHULUAN

Perkembangan teknologi digital dan media sosial telah meningkatkan jumlah data teks yang dihasilkan pengguna internet setiap harinya. Salah satu platform media sosial dengan tingkat interaksi tinggi adalah YouTube, di mana pengguna dapat memberikan komentar terhadap video yang mereka tonton. Komentar tersebut mengandung opini, kritik, apresiasi, serta berbagai bentuk respons publik terhadap suatu konten digital. Data komentar pada media sosial menjadi sumber informasi yang penting untuk dianalisis menggunakan pendekatan *text mining* dan *Natural Language Processing* (NLP) (Alpiana et al., 2024; Nanayakkara & Thennakoon, 2024).

Analisis komentar media sosial telah banyak digunakan untuk memahami opini publik terhadap berbagai topik seperti hiburan, pemasaran digital, politik, dan isu sosial. Penelitian Lee dan Nguyen (2023) menunjukkan bahwa komentar YouTube dapat dimanfaatkan untuk mengidentifikasi topik pembahasan serta sentimen pengguna terhadap konten *sustainable fashion*. Selain itu, komentar YouTube juga digunakan dalam analisis sentimen menggunakan pendekatan *deep learning* untuk memahami respons publik terhadap berbagai isu sosial dan hiburan (A. Aiswarya & H. Rajeev, 2024; Giri et al., 2024). Hal tersebut menunjukkan bahwa komentar

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

YouTube memiliki potensi besar sebagai sumber data untuk memahami opini dan pola pembahasan masyarakat secara otomatis.

Salah satu pendekatan yang banyak digunakan dalam analisis teks media sosial adalah *topic modeling*. Metode ini digunakan untuk menemukan pola topik tersembunyi dari kumpulan dokumen teks berdasarkan distribusi kata pada dokumen (Egger & Yu, 2022). Beberapa metode *topic modeling* yang umum digunakan antara lain BERTopic, *Non-negative Matrix Factorization* (NMF), Top2Vec, dan *Latent Dirichlet Allocation* (LDA) (Egger & Yu, 2022). Penelitian Egger dan Yu (2022) membandingkan LDA, NMF, Top2Vec, dan BERTopic pada data Twitter dan menunjukkan bahwa metode berbasis *embedding* seperti BERTopic mampu menghasilkan topik yang lebih kontekstual. Namun, metode tersebut membutuhkan sumber daya komputasi yang lebih tinggi dan proses interpretasi topik yang lebih kompleks, terutama pada teks pendek media sosial. George dan Sumathy (2023) juga mengembangkan pendekatan berbasis BERT untuk meningkatkan kualitas *topic modeling*, tetapi pendekatan tersebut memerlukan proses pelatihan model yang lebih kompleks dibandingkan metode probabilistik konvensional.

Metode *Latent Dirichlet Allocation* merupakan algoritma *topic modeling* berbasis probabilistik yang digunakan untuk menemukan topik tersembunyi pada kumpulan dokumen teks. Penelitian Alpiana et al. (2024) menunjukkan bahwa metode LDA mampu mengidentifikasi topik dominan pada ulasan media sosial YouTube melalui evaluasi *coherence score*. Penelitian Wyskwarski (2024) juga menunjukkan bahwa LDA efektif digunakan untuk menganalisis komentar YouTube terkait kendaraan listrik dan menghasilkan topik yang mudah diinterpretasikan. Selain itu, Gomes dan Attux (2023) menyatakan bahwa metode LDA masih relevan dalam analisis media sosial karena memiliki implementasi yang lebih sederhana, stabil, dan mudah dipahami dibandingkan beberapa metode *topic modeling* modern lainnya. Oleh karena itu, LDA masih dianggap efektif untuk menganalisis komentar media sosial yang bersifat tidak terstruktur dan berjumlah besar.

Video klip *Lighter* sebagai lagu tema FIFA World Cup 2026 memperoleh perhatian besar dari pengguna internet melalui platform YouTube dan menghasilkan banyak komentar yang membahas kualitas lagu, atmosfer sepak bola, visual video, serta representasi FIFA World Cup itu sendiri. Banyaknya komentar yang tersedia membuat proses analisis secara manual menjadi kurang efektif sehingga diperlukan pendekatan otomatis menggunakan teknik *Information Retrieval* (IR) dan NLP. Pengambilan data komentar dilakukan menggunakan teknik *web scraping* melalui platform Apify berdasarkan *top comments*. Teknik *web scraping* banyak digunakan dalam penelitian media sosial karena mampu mengambil data komentar secara otomatis dalam jumlah besar (S et al., 2024; Su et al., 2024). Setelah proses pengambilan data dilakukan, tahapan *preprocessing* seperti *cleaning text*, *tokenization*, *stopword removal*, dan pembentukan *bigram* diterapkan untuk meningkatkan kualitas data sebelum dilakukan proses *topic modeling*.

Beberapa penelitian sebelumnya telah membahas *topic modeling* pada komentar YouTube dan media sosial lainnya (Alpiana et al., 2024; Bergamini Gomes & Attux, 2023; Egger & Yu, 2022; Wyskwarski, 2024). Namun, penelitian yang secara khusus menganalisis topik komentar YouTube pada video klip *Lighter* sebagai lagu tema FIFA World Cup 2026 masih sangat terbatas. Selain itu, sebagian besar penelitian sebelumnya berfokus pada analisis sentimen atau isu politik dan pemasaran digital, sedangkan analisis topik pada konten hiburan olahraga internasional masih belum banyak dibahas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis topik dominan pada komentar YouTube video klip *Lighter* menggunakan metode *Latent Dirichlet Allocation*. Hasil penelitian diharapkan dapat membantu memahami pola opini publik terhadap lagu tema FIFA World Cup 2026 melalui analisis komentar YouTube secara otomatis.

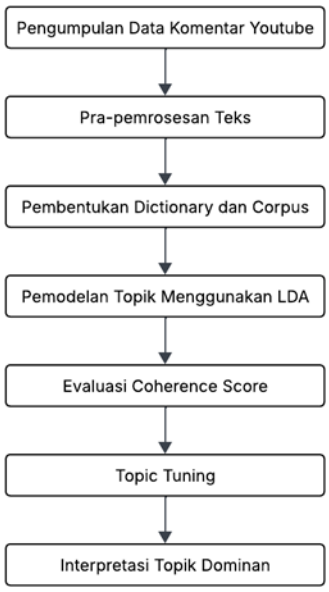
2. METODE PENELITIAN

2.1 Desain dan Alur Penelitian

Penelitian ini bertujuan untuk mengidentifikasi pola pembahasan yang berkembang pada komentar pengguna YouTube terhadap video musik *Lighter* yang digunakan sebagai lagu resmi FIFA World Cup 2026. Proses penelitian diawali dengan pengambilan data komentar, kemudian dilanjutkan dengan pengolahan dan normalisasi teks agar data siap dianalisis. Selanjutnya, data direpresentasikan ke dalam format yang dapat diproses oleh model *Latent Dirichlet Allocation* (LDA) untuk memperoleh topik-topik yang muncul dalam komentar

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

pengguna. Kualitas model dinilai menggunakan *coherence score,* sedangkan pemilihan jumlah topik dilakukan melalui perbandingan beberapa konfigurasi model. Tahap akhir penelitian berupa analisis dan penafsiran terhadap topik yang terbentuk untuk memahami kecenderungan diskusi pengguna. Rangkaian tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Flowchart Alur Penelitian

2.2 Pengumpulan Data

Data yang dianalisis dalam penelitian ini berupa komentar pengguna pada video musik Lighter yang dipublikasikan melalui platform YouTube sebagai lagu resmi FIFA World Cup 2026. Pengambilan data dilakukan secara otomatis dengan bantuan layanan Apify sehingga proses pengumpulan komentar dapat dilakukan secara efisien. Sebanyak 398 komentar berbahasa Inggris berhasil diperoleh dan digunakan sebagai sumber data utama. Komentar tersebut mencerminkan berbagai pandangan pengguna mengenai kualitas lagu, nuansa sepak bola yang ditampilkan, serta aspek lain yang berkaitan dengan video musik FIFA.

2.3 Pra-pemrosesan Teks

Untuk meningkatkan kualitas data sebelum dilakukan pemodelan topik, seluruh komentar terlebih dahulu diproses melalui tahap *preprocessing*. Tahapan ini bertujuan mengurangi *noise* dan menyeragamkan bentuk teks agar lebih mudah dianalisis. Proses yang dilakukan meliputi konversi huruf ke format *lowercase*, penghapusan URL, angka, tanda baca, emoji, dan karakter yang tidak diperlukan, pemisahan teks menjadi token kata, penghilangan *stopword*, pembentukan *bigram*, serta *lemmatization* untuk mengubah kata ke bentuk dasarnya. Hasil dari tahapan ini berupa kumpulan teks yang lebih terstruktur dan siap digunakan pada proses pemodelan topik.

2.4 Pembentukan Dictionary dan Corpus

Data yang telah melalui tahap pra-pemrosesan selanjutnya direpresentasikan ke dalam format yang dapat diproses oleh model LDA. Proses ini dilakukan menggunakan pustaka Gensim dengan membentuk *dictionary* dan *corpus*. *Dictionary* berfungsi sebagai pemetaan antara kata unik dan identitas numeriknya, sedangkan *corpus* menyimpan representasi dokumen

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

berdasarkan frekuensi kemunculan kata dalam setiap komentar. Representasi tersebut digunakan sebagai dasar dalam proses pembentukan model topik.

2.5 Pemodelan Topik Menggunakan LDA

Metode Latent Dirichlet Allocation (LDA) digunakan untuk menemukan topik tersembunyi dari kumpulan komentar YouTube secara otomatis. LDA merupakan metode topic modeling berbasis probabilistik yang mengasumsikan bahwa setiap dokumen terdiri atas beberapa topik dan setiap topik direpresentasikan oleh distribusi kata tertentu. Probabilitas suatu kata w pada dokumen d dapat direpresentasikan menggunakan Persamaan 1.

$$P(w) = \sum_{z=1}^{K} P(w \mid z)P(z \mid d) \quad (1)$$

Dengan:

1. P(w)= probabilitas kemunculan kata
2. P(w | z) = probabilitas kata pada topik tertentu
3. P(z | d) = probabilitas topik pada dokumen
4. K = jumlah topik

Dalam penelitian ini, model LDA dibangun menggunakan pustaka Gensim dengan beberapa variasi jumlah topik untuk memperoleh model terbaik. Setiap topik direpresentasikan melalui kumpulan kata dengan probabilitas tertinggi.

2.6 Evaluasi Model Menggunakan Coherence Score

Kualitas topik yang dihasilkan dievaluasi menggunakan nilai *coherence score*. Metrik ini digunakan untuk menilai tingkat keterkaitan antar kata dalam suatu topik sehingga dapat diketahui sejauh mana topik yang terbentuk memiliki makna yang konsisten.
Pada penelitian ini digunakan metode *c_v coherence* yang tersedia pada pustaka Gensim. Evaluasi dilakukan terhadap beberapa variasi jumlah topik untuk memperoleh model yang menghasilkan topik paling representatif. Nilai coherence score digunakan sebagai dasar dalam menentukan jumlah topik terbaik pada proses topic tuning.

2.7 Topic Tuning dan Interpretasi Topik

Proses topic tuning dilakukan dengan membandingkan beberapa variasi jumlah topik berdasarkan nilai coherence score yang diperoleh. Model dengan nilai coherence terbaik dipilih sebagai model akhir penelitian.
Setelah model terbaik diperoleh, dilakukan interpretasi topik dengan menganalisis kata-kata dominan pada setiap topik. Interpretasi ini bertujuan untuk mengidentifikasi tema utama yang paling sering dibahas pengguna dalam komentar video klip Lighter sebagai lagu tema FIFA World Cup 2026.

3. ANALISA DAN PEMBAHASAN

Hasil penelitian pada bagian ini berfokus pada identifikasi pola pembahasan yang muncul dalam komentar pengguna YouTube terhadap video musik *Lighter*. Data yang dianalisis berasal dari 398 komentar berbahasa Inggris yang berhasil dikumpulkan secara otomatis melalui layanan *web scraping*. Setelah melewati tahapan pembersihan dan normalisasi teks, data kemudian diproses menggunakan metode *Latent Dirichlet Allocation* untuk menemukan tema-tema dominan yang muncul dalam diskusi pengguna.

3.1 Hasil Topic Modeling

Berdasarkan proses pemodelan menggunakan metode LDA, diperoleh tiga topik utama yang merepresentasikan pola pembahasan dominan pada komentar pengguna YouTube. Hasil topic modeling ditunjukkan pada Tabel 1.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

Tabel 1. Hasil Topic Modeling Menggunakan LDA

Topik	Kata Dominan	Interpretasi
Topik 0	like, better, dont, know, sounds, american	Opini pengguna terhadap kualitas musik
Topik 1	thats, thick, football, doesnt, american	Kritik terhadap kesesuaian lagu dengan tema sepak bola
Topik 2	music, video, fifa, sounds	Diskusi umum mengenai video musik FIFA

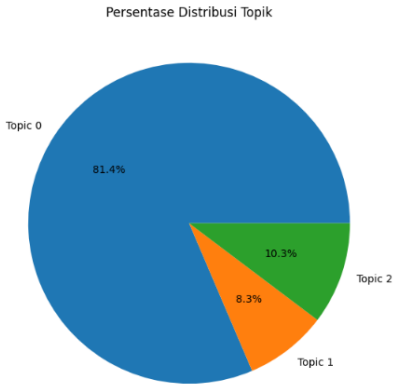
Topik 0 merupakan topik yang paling dominan dan berisi komentar mengenai kualitas lagu serta tanggapan pengguna terhadap musik yang digunakan. Topik 1 menunjukkan adanya kritik terhadap kesesuaian lagu dengan nuansa FIFA World Cup. Sementara itu, Topik 2 berisi pembahasan umum mengenai video musik FIFA dan interaksi pengguna pada kolom komentar. Beberapa komentar yang mewakili masing-masing topik ditunjukkan pada Tabel 2.

Tabel 2. Contoh Komentar Berdasarkan Topik Dominan

Topik	Contoh Komentar	Interpretasi
Topik 0	“you guys are just hating like the song is so much better at full volume”	Opini positif terhadap kualitas lagu
Topik 0	“I’ma play this at my funeral so that people are jealous when I’m dead”	Respon humor dan opini
Topik 1	“this doesn’t sound like football music”	Kritik terhadap kesesuaian lagu dengan tema FIFA
Topik 1	“american music doesn’t fit world cup vibes”	Kritik terhadap nuansa musik yang dianggap tidak sesuai
Topik 2	“bro is stealing the likes from the video”	Diskusi umum mengenai video dan interaksi pengguna
Topik 2	“music video fifa sounds amazing”	Pembahasan umum terkait video musik FIFA

3.2 Distribusi dan Visualisasi Topik

Distribusi topik hasil pemodelan LDA ditunjukkan pada Gambar 2.

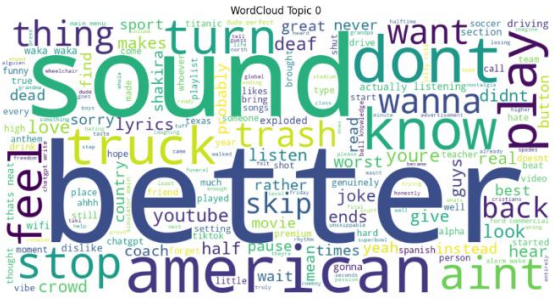


Gambar 2. Distribusi Topik Komentar

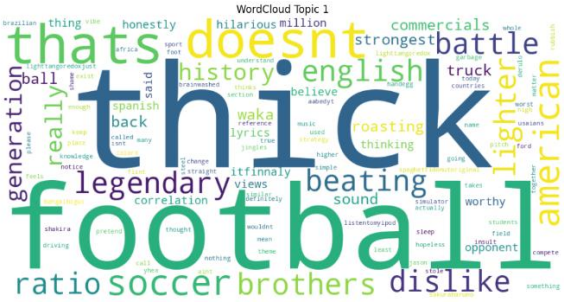
Berdasarkan hasil distribusi topik, Topik 0 memiliki persentase tertinggi sebesar 81,4%, sedangkan Topik 1 dan Topik 2 masing-masing sebesar 8,3% dan 10,3%. Hasil ini menunjukkan bahwa sebagian besar pengguna lebih banyak memberikan opini terkait kualitas musik dibandingkan pembahasan tema sepak bola maupun video FIFA secara umum.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 1, Tahun 2026
ISSN 3025-0919 (media online)
Hal 303-309

Visualisasi WordCloud digunakan untuk menampilkan kata-kata yang paling sering muncul pada setiap topik hasil pemodelan. Hasil visualisasi WordCloud ditunjukkan pada Gambar 3, Gambar 4 dan Gambar 5.



Gambar 3. WordCloud Topic 0



Gambar 4. WordCloud Topic 1

Gambar 5. WordCloud Topic 2

Pada Topik 0 terlihat dominasi kata seperti music, sounds, dan like yang menunjukkan fokus pengguna terhadap kualitas lagu. Pada Topik 1 muncul kata football dan american yang menggambarkan kritik terhadap kesesuaian lagu dengan atmosfer FIFA World Cup. Sementara itu, Topik 2 didominasi kata video dan fifa yang menunjukkan pembahasan umum terkait video musik FIFA.

3.3 Evaluasi Model dan Pembahasan

Evaluasi model dilakukan menggunakan coherence score untuk mengukur kualitas topik yang dihasilkan. Berdasarkan hasil pengujian, model memperoleh nilai coherence score sebesar 0,466. Nilai tersebut menunjukkan bahwa topik yang dihasilkan memiliki keterkaitan kata yang cukup baik dan dapat diinterpretasikan dengan jelas.

Selain itu, pengujian jumlah topik dari 2 hingga 10 menunjukkan bahwa nilai coherence score relatif stabil. Oleh karena itu, dipilih tiga topik sebagai model akhir karena menghasilkan interpretasi yang paling sederhana dan representatif terhadap data komentar yang dianalisis.

Hasil penelitian ini sejalan dengan penelitian Alpiana et al. yang menunjukkan bahwa metode LDA mampu mengidentifikasi topik dominan pada komentar YouTube secara efektif. Temuan

penelitian ini juga mendukung penelitian Wyskwarski yang menyatakan bahwa LDA masih relevan digunakan pada data media sosial karena menghasilkan topik yang mudah dipahami dan diinterpretasikan. Dengan demikian, metode LDA terbukti mampu memberikan gambaran pola opini publik terhadap lagu tema FIFA World Cup 2026 melalui analisis komentar YouTube.

4. KESIMPULAN

Penelitian ini berhasil menerapkan metode Latent Dirichlet Allocation (LDA) untuk menganalisis topik komentar YouTube pada video klip Lighter sebagai lagu tema FIFA World Cup 2026. Berdasarkan hasil analisis terhadap 398 komentar berbahasa Inggris, diperoleh tiga topik utama yang membahas opini pengguna terhadap kualitas musik, kritik mengenai kesesuaian lagu dengan nuansa sepak bola, serta diskusi umum terkait video musik FIFA. Hasil distribusi topik menunjukkan bahwa sebagian besar komentar didominasi oleh pembahasan mengenai kualitas lagu dengan persentase sebesar 81,4%. Selain itu, evaluasi model menggunakan coherence score menghasilkan nilai sebesar 0,466 yang menunjukkan bahwa topik yang dihasilkan cukup koheren dan dapat diinterpretasikan dengan baik. Hasil penelitian ini menunjukkan bahwa metode LDA efektif digunakan untuk mengidentifikasi pola pembahasan dan opini publik pada komentar YouTube berbasis teks pendek. Untuk penelitian selanjutnya, analisis dapat dikembangkan dengan membandingkan metode LDA dengan pendekatan topic modeling lain seperti BERTopic atau Top2Vec, serta mengombinasikannya dengan analisis sentimen untuk memperoleh informasi yang lebih komprehensif mengenai respons pengguna.

REFERENCES

Aiswarya, A., & Rajeev, H. (2024). Youtube comment sentimental analysis. *Indian Journal of Data Mining, 4* (1), 5–8. https://doi.org/10.54105/ijdm.A1633.04010524

Alpiana, V., Salam, A., Alzami, F., Rizqa, I., & Aqmala, D. (2024). Analisis topic-modelling menggunakan Latent Dirichlet Allocation (LDA) pada ulasan sosial media YouTube. *Jurnal Media Informatika Budidarma, 8* (1), 332–341. https://doi.org/10.30865/mib.v8i1.7127

Bergamini Gomes, G. B., & Attux, R. (2023). Contributions to social media analysis based on topic modelling. *Anais do XI Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2023)*, 113–120. https://doi.org/10.5753/kdmile.2023.231795

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology, 7*, Article 886498. https://doi.org/10.3389/fsoc.2022.886498

Giri, R., Sirsath, M., & Kanakia, H. T. (2024). Youtube comments sentiment analysis. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1–4). IEEE. https://doi.org/10.1109/I2CT61223.2024.10544022

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, 17* (1), 40–47. https://doi.org/10.4038/icter.v17i1.7276

S, D., Deep, A., Mishra, A., Mishra, A. C., & Gangoor, S. A. (2024). Youtube comments scraping and sentiment analysis. *International Journal of Scientific Research in Engineering and Management, 8* (4), 1–5. https://doi.org/10.55041/IJSREM32103

Su, L. Y.-F., Chen, T., Ng, Y. M. M., Gong, Z., & Wang, Y.-C. (2024). Integrating human insights into text analysis: Semi-supervised topic modeling of emerging food-technology businesses' brand communication on social media. *Social Science Computer Review, 42* (2), 416–437. https://doi.org/10.1177/08944393231184532

Wyskwarski, M. (2024). Uncovering topics in YouTube comments on electric vehicles using Latent Dirichlet Allocation. *Scientific Papers of Silesian University of Technology: Organization and Management Series, 210*, 671–686. https://doi.org/10.29119/1641-3466.2024.210.44