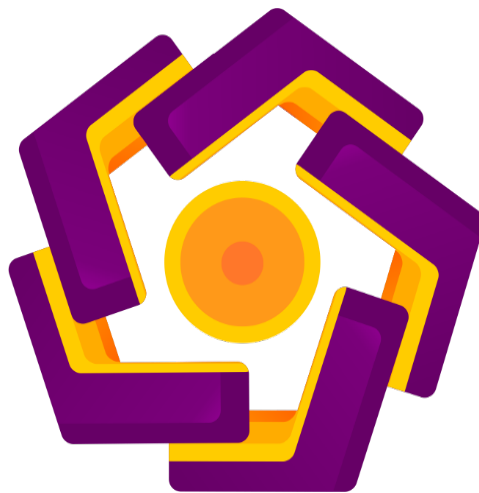


**PEMETAAN ISU JUDI ONLINE DI INDONESIA MELALUI  
PEMODELAN TOPIK MENGGUNAKAN LDA DAN NMF**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**OKTAVIANUS MARIO**

**21.11.4138**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2026**

**PEMETAAN ISU JUDI ONLINE DI INDONESIA MELALUI  
PEMODELAN TOPIK MENGGUNAKAN LDA DAN NMF**

**SKRIPSI**

untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**OKTAVIANUS MARIO**

**21.11.4138**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2026**

**HALAMAN PERSETUJUAN**

**SKRIPSI**

**Pemetaan Isu Judi Online di Indonesia melalui Pemodelan Topik  
Menggunakan LDA dan NMF**

yang disusun dan diajukan oleh

**Oktavianus Marlo**

**21.11.4138**

telah disetujui oleh Dosen Pembimbing Skripsi  
pada tanggal 18 Februari 2026

**Dosen Pembimbing,**



**Anna Baita S.Kom., M.Kom.**

**NIK. 190302290**

## HALAMAN PENGESAHAN

### SKRIPSI

**Pemetaan Isu Judi Online di Indonesia melalui Pemodelan Topik  
Menggunakan LDA dan NMF**

yang disusun dan diajukan oleh

**Oktavianus Mario**

**21.11.4138**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 18 Februari 2026

#### Susunan Dewan Penguji

**Nama Penguji**

**Tanda Tangan**

**Ferian Fauzi Abdulloh, S.Kom., M.Kom.**

**NIK. 190302276**

**Dr. Hartatik, S.T., M.Cs.**

**NIK. 190302232**

**Anna Baita S.Kom., M.Kom.**

**NIK. 190302290**

Skripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 18 Februari 2026

**DEKAN FAKULTAS ILMU KOMPUTER**



**Prof. Dr. Kusriani, M.Kom.**

**NIK. 190302106**

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Oktavianus Mario  
NIM : 21.11.4138

Menyatakan bahwa Skripsi dengan judul berikut:

**Pemetaan Isu Judi Online di Indonesia melalui Pemodelan Topik  
Menggunakan LDA dan NMF**  
Dosen Pembimbing : Anna Baita, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 18 Februari 2026

Yang Menyatakan,



Oktavianus Mario

## KATA PENGANTAR

Puji syukur saya panjatkan ke hadirat Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya, sehingga saya dapat menyelesaikan skripsi sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Informatika Universitas Amikom Yogyakarta.

Penulisan skripsi ini tidak lepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini, saya menyampaikan ucapan terima kasih kepada:

1. Ibu Dosen Pembimbing, **Anna Baita, M.Kom**, yang telah memberikan bimbingan, arahan, serta motivasi selama proses penyusunan skripsi ini.
2. Tim Dosen Penguji yang telah memberikan kritik, saran, dan masukan yang sangat berharga demi perbaikan karya tulis ini.
3. Kedua Orang tua tercinta, Bapak Fransiskus Hendrikus dan Ibu Maria Asnita Yeti yang selalu memberikan dukungan moril, materil, serta doa yang tak pernah putus.
4. Kekasih peneliti Sydney Ilma Zees yang selalu menemani saat mengerjakan, membantu dan memberi masukan, serta memberi semangat dan doa agar skripsi ini segera selesai.
5. Sahabat peneliti Alexsander Geno, Andrias Setiawan, dan Akmal Fauzan yang menemani dan membantu dalam penyelesaian skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi penyempurnaan karya tulis ini di masa yang akan datang.

Akhirnya, semoga skripsi ini dapat memberikan manfaat bagi penulis khususnya dan pembaca pada umumnya.

Yogyakarta,



Oktavianus Mario

## DAFTAR ISI

|                                                         |      |
|---------------------------------------------------------|------|
| HALAMAN JUDUL .....                                     | i    |
| HALAMAN PERSETUJUAN.....                                | ii   |
| HALAMAN PENGESAHAN .....                                | iii  |
| HALAMAN PERNYATAAN KEASLIAN SKRIPSI .....               | iv   |
| KATA PENGANTAR .....                                    | v    |
| DAFTAR ISI.....                                         | vi   |
| DAFTAR TABEL.....                                       | ix   |
| DAFTAR GAMBAR .....                                     | x    |
| DAFTAR LAMPIRAN.....                                    | xi   |
| DAFTAR LAMBANG DAN SINGKATAN .....                      | xii  |
| DAFTAR ISTILAH .....                                    | xiv  |
| INTISARI .....                                          | xxiv |
| ABSTRACT.....                                           | xxv  |
| BAB I PENDAHULUAN.....                                  | 1    |
| 1.1 Latar Belakang .....                                | 1    |
| 1.2 Rumusan Masalah .....                               | 2    |
| 1.3 Batasan Masalah .....                               | 2    |
| 1.4 Tujuan Penelitian .....                             | 3    |
| 1.5 Manfaat Penelitian .....                            | 3    |
| 1.6 Sistematika Penulisan .....                         | 4    |
| BAB II TINJAUAN PUSTAKA .....                           | 6    |
| 2.1 Studi Literatur .....                               | 6    |
| 2.2 Dasar Teori.....                                    | 16   |
| 2.2.1 Judi Online .....                                 | 16   |
| 2.2.2 YouTube sebagai Sumber Data Komentar Publik ..... | 16   |
| 2.2.3 Text Mining .....                                 | 17   |
| 2.2.4 Data Collection .....                             | 17   |
| 2.2.5 Exploratory Data Analysis .....                   | 17   |
| 2.2.6 Text Preprocessing .....                          | 18   |
| 2.2.7 Representasi Teks .....                           | 18   |
| 2.2.8 Analisis Teks.....                                | 19   |

|                                   |                                                       |    |
|-----------------------------------|-------------------------------------------------------|----|
| 2.2.9                             | Topic Modeling.....                                   | 20 |
| 2.2.10                            | Latent Dirichlet Allocation (LDA) .....               | 20 |
| 2.2.11                            | Nonnegative Matrix Factorization (NMF).....           | 22 |
| 2.2.12                            | Evaluasi.....                                         | 23 |
| 2.2.13                            | Evaluasi Topic Modeling .....                         | 23 |
| BAB III METODE PENELITIAN .....   |                                                       | 25 |
| 3.1                               | Alur Penelitian .....                                 | 25 |
| 3.2                               | Tahapan Penelitian.....                               | 26 |
| 3.2.1                             | Pengumpulan Data .....                                | 26 |
| 3.2.2                             | Validasi Data.....                                    | 26 |
| 3.2.3                             | Analisis Eksploratif Data Awal .....                  | 26 |
| 3.2.4                             | Pra-pemrosesan Teks .....                             | 27 |
| 3.2.5                             | Analisis Eksploratif Data Setelah Pra-pemrosesan..... | 27 |
| 3.2.6                             | Representasi Teks .....                               | 28 |
| 3.2.7                             | Pemodelan Topik .....                                 | 28 |
| 3.2.8                             | Evaluasi Model .....                                  | 29 |
| 3.2.9                             | Interpretasi Topik.....                               | 29 |
| 3.2.10                            | Visualisasi dan Analisis Hasil.....                   | 30 |
| BAB IV HASIL DAN PEMBAHASAN ..... |                                                       | 31 |
| 4.1                               | Pengumpulan Data .....                                | 31 |
| 4.2                               | Validasi Data.....                                    | 32 |
| 4.3                               | Analisis Eksploratif Data Awal .....                  | 34 |
| 4.4                               | Pra-pemrosesan Teks .....                             | 36 |
| 4.4.1                             | Case Folding .....                                    | 36 |
| 4.4.2                             | Pembersihan Teks .....                                | 37 |
| 4.4.3                             | Normalisasi Kata.....                                 | 38 |
| 4.4.4                             | Tokenisasi .....                                      | 39 |
| 4.4.5                             | Penghapusan Stopword.....                             | 39 |
| 4.4.6                             | Stemming .....                                        | 39 |
| 4.4.7                             | Pembentukan Bigram.....                               | 40 |
| 4.4.8                             | Penyaringan Dokumen Pendek .....                      | 41 |
| 4.4.9                             | Hasil Keseluruhan Pra-pemrosesan .....                | 41 |
| 4.5                               | Analisis Eksploratif Data Setelah Pra-pemrosesan..... | 43 |
| 4.6                               | Representasi Teks .....                               | 44 |

|                |                                                |    |
|----------------|------------------------------------------------|----|
| 4.6.1          | Bag-of-Words untuk LDA .....                   | 45 |
| 4.6.2          | TF-IDF untuk NMF .....                         | 46 |
| 4.7            | Pemodelan Topik .....                          | 47 |
| 4.7.1          | Eksperimen Penentuan Jumlah Topik Optimal..... | 47 |
| 4.7.2          | Hasil LDA dan NMF Topik Optimal.....           | 48 |
| 4.7.3          | Interpretasi Hasil Awal .....                  | 49 |
| 4.7.3.1        | Interpretasi Topik LDA (7 Topik Awal).....     | 49 |
| 4.7.4          | Visualisasi Hasil Awal.....                    | 52 |
| 4.7.4.1        | Visualisasi Topik LDA Awal.....                | 52 |
| 4.8            | Evaluasi.....                                  | 54 |
| 4.9            | Interpretasi Topik.....                        | 55 |
| 4.9.1          | Analisis Hasil Interpretasi Awal .....         | 56 |
| 4.9.2          | Penyaringan dan Penyempurnaan Topik .....      | 57 |
| 4.9.3          | Hasil Akhir Interpretasi Topik LDA.....        | 57 |
| 4.9.4          | Hasil Akhir Interpretasi Topik NMF .....       | 58 |
| 4.9.5          | Perbandingan dan Sintesis Hasil.....           | 59 |
| BAB V          | PENUTUP .....                                  | 63 |
| 5.1            | Kesimpulan .....                               | 63 |
| 5.2            | Saran .....                                    | 64 |
| Daftar Pustaka | .....                                          | 65 |
| LAMPIRAN       | .....                                          | 71 |

## DAFTAR TABEL

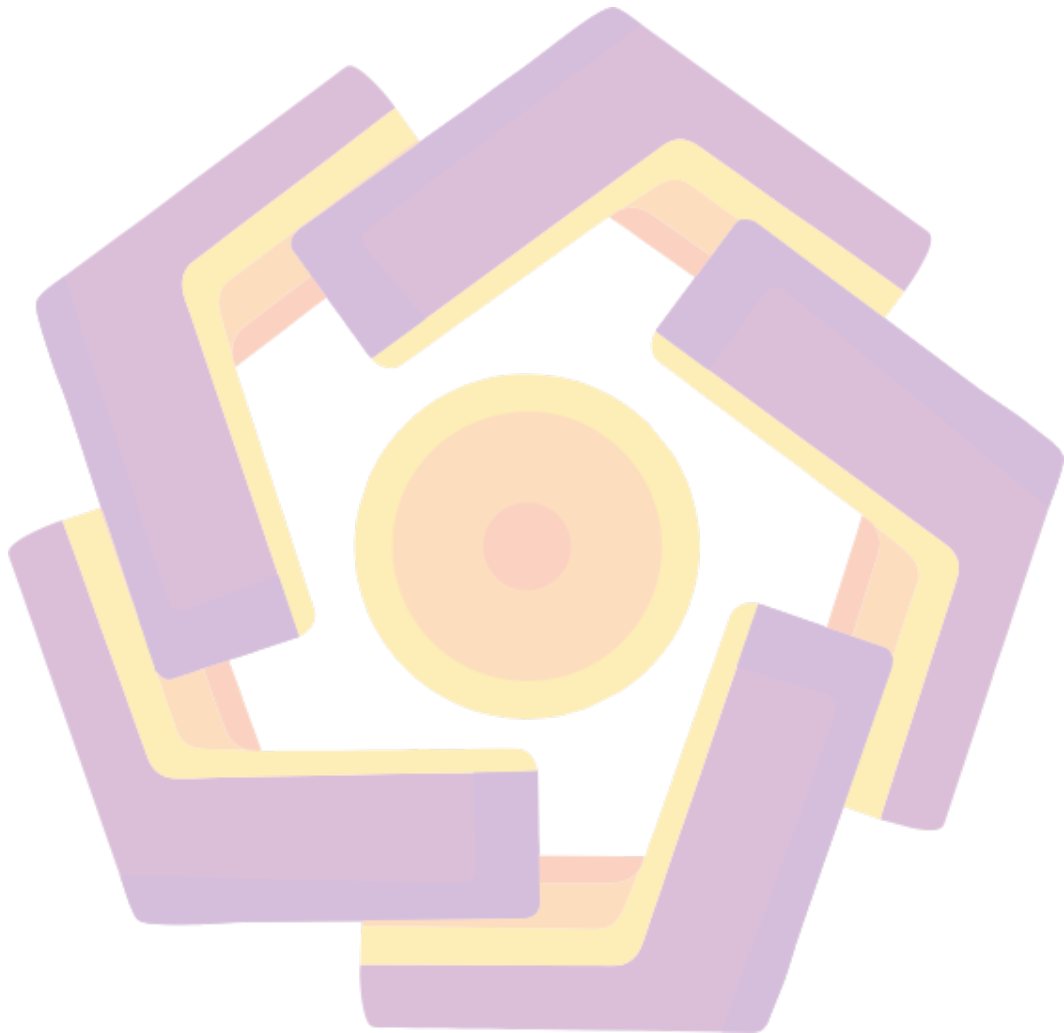
|                                                                     |    |
|---------------------------------------------------------------------|----|
| Tabel 2. 1. Keaslian Penelitian .....                               | 8  |
| Tabel 4. 1. Distribusi Komentar per Video.....                      | 31 |
| Tabel 4. 2. Hasil Validasi Data .....                               | 33 |
| Tabel 4. 3. Dua Puluh Kata Paling Sering Muncul (Data Mentah) ..... | 34 |
| Tabel 4. 4. Contoh Transformasi Pembersihan Teks .....              | 37 |
| Tabel 4. 5. Contoh Normalisasi Kata.....                            | 38 |
| Tabel 4. 6. Contoh Bigram yang Terbentuk .....                      | 40 |
| Tabel 4. 7. Statistik Panjang Dokumen Setelah Pra-pemrosesan .....  | 42 |
| Tabel 4. 8. Contoh Transformasi Teks Melalui Pra-pemrosesan.....    | 42 |
| Tabel 4. 9. Parameter dan Hasil Bag-of-Words .....                  | 45 |
| Tabel 4. 10. Parameter dan Hasil TF-IDF.....                        | 46 |
| Tabel 4. 11. Konfigurasi Parameter Pencarian Topik Optimal.....     | 47 |
| Tabel 4. 12. Hasil Eksperimen Pencarian Topik Optimal .....         | 48 |
| Tabel 4. 13. Interpretasi Topik LDA (Hasil Awal - 7 Topik).....     | 49 |
| Tabel 4. 14. Interpretasi Topik NMF (Hasil Awal - 8 Topik) .....    | 50 |
| Tabel 4. 15. Ringkasan hasil evaluasi model .....                   | 55 |
| Tabel 4. 16. Temuan Masalah pada Topik LDA .....                    | 56 |
| Tabel 4. 18. Distribusi dan Interpretasi Topik Model LDA .....      | 58 |
| Tabel 4. 19. Distribusi dan Interpretasi Topik Model NMF .....      | 59 |

## DAFTAR GAMBAR

|                                                                                           |    |
|-------------------------------------------------------------------------------------------|----|
| Gambar 2. 1. Logo Youtube .....                                                           | 16 |
| Gambar 2. 2. Grafik Model LDA.....                                                        | 21 |
| Gambar 3. 1. Alur Penelitian .....                                                        | 25 |
| Gambar 4. 1. Distribusi Panjang Teks Sebelum dan Sesudah Validasi.....                    | 33 |
| Gambar 4. 2. Histogram Distribusi Panjang Komentar (Data Mentah).....                     | 36 |
| Gambar 4. 3. <i>Word Cloud</i> Kata-kata Paling Sering Setelah <i>Preprocessing</i> ..... | 44 |
| Gambar 4. 4. <i>Word Cloud</i> Topik LDA (Hasil Awal - 7 Topik).....                      | 52 |
| Gambar 4. 5. <i>Word Cloud</i> Topik NMF (Hasil Awal - 8 Topik).....                      | 53 |
| Gambar 4. 6. <i>Word Cloud</i> Topik LDA .....                                            | 60 |
| Gambar 4. 7. <i>Word Cloud</i> Topik NMF .....                                            | 61 |
| Gambar 4. 8. Distribusi Prevalensi Topik.....                                             | 62 |

## DAFTAR LAMPIRAN

|                              |    |
|------------------------------|----|
| Lampiran 1. Link Github..... | 71 |
|------------------------------|----|

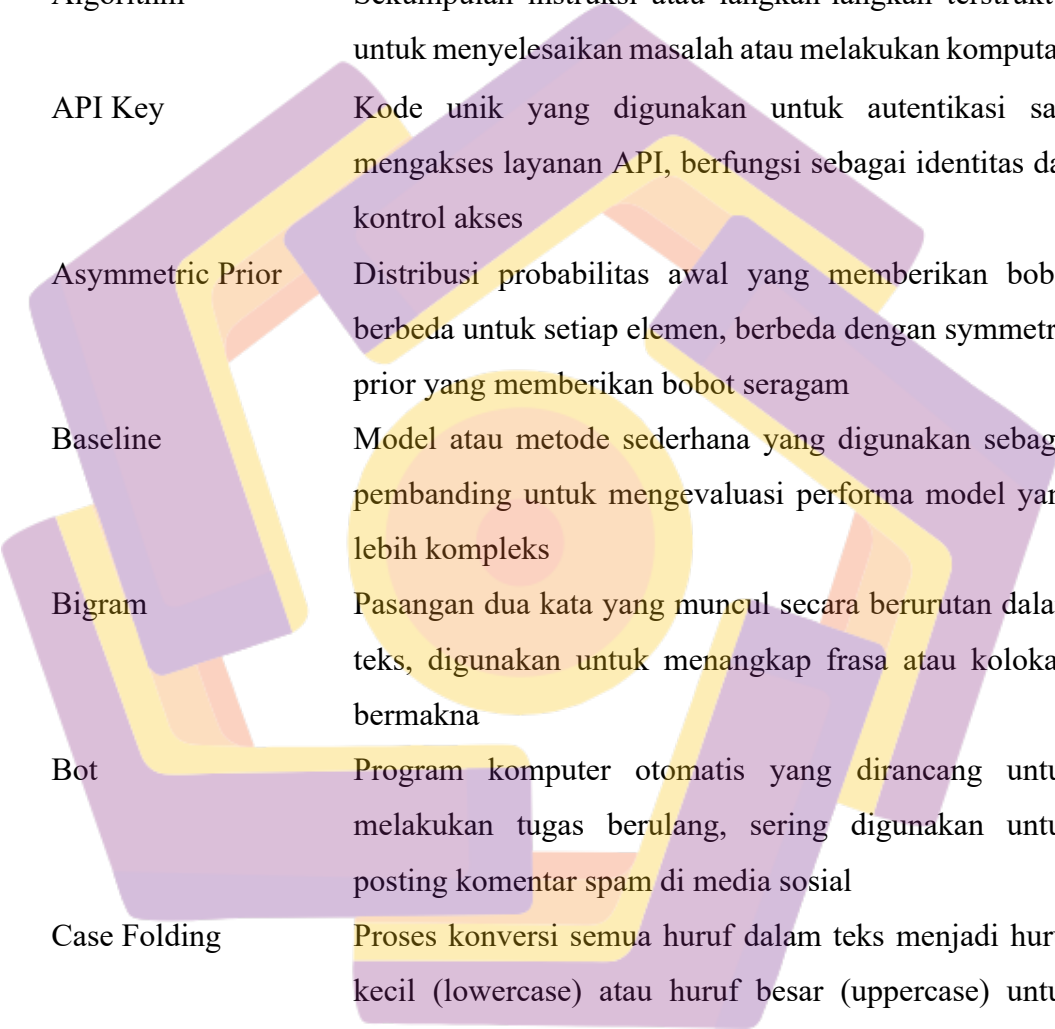


## DAFTAR LAMBANG DAN SINGKATAN

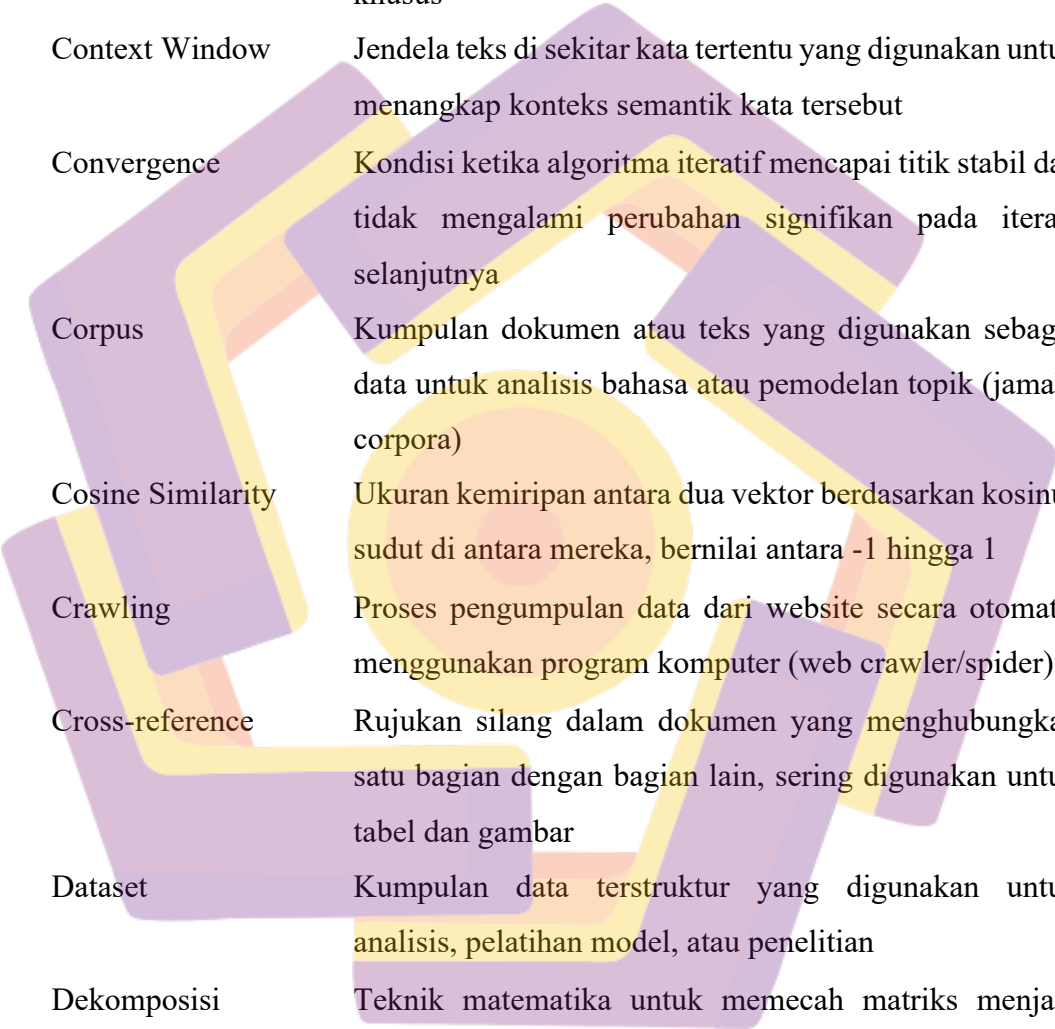
|                  |                                                                                             |
|------------------|---------------------------------------------------------------------------------------------|
| $\alpha$ (alpha) | Parameter prior distribusi topik per dokumen dalam LDA                                      |
| $\beta$ (beta)   | Parameter distribusi kata per topik dalam LDA                                               |
| $\eta$ (eta)     | Parameter prior distribusi kata per topik dalam LDA                                         |
| $\theta$ (theta) | Distribusi topik dalam dokumen                                                              |
| $\phi$ (phi)     | Distribusi kata dalam topik                                                                 |
| API              | Application Programming Interface                                                           |
| BoW              | Bag-of-Words                                                                                |
| CD               | Coordinate Descent                                                                          |
| CPU              | Central Processing Unit                                                                     |
| CSV              | Comma-Separated Values                                                                      |
| $C_v$            | Coherence measure menggunakan normalized pointwise mutual information dan cosine similarity |
| DF               | Document Frequency                                                                          |
| EDA              | Exploratory Data Analysis                                                                   |
| GPU              | Graphics Processing Unit                                                                    |
| H                | Matriks topik-kata dalam NMF                                                                |
| HTML             | HyperText Markup Language                                                                   |
| IDF              | Inverse Document Frequency                                                                  |
| JSON             | JavaScript Object Notation                                                                  |
| K                | Jumlah topik dalam model                                                                    |
| LDA              | Latent Dirichlet Allocation                                                                 |
| MCP              | Model Context Protocol                                                                      |
| ML               | Machine Learning                                                                            |
| N                | Jumlah total dokumen dalam korpus                                                           |
| NLP              | Natural Language Processing                                                                 |
| NMF              | Non-Negative Matrix Factorization                                                           |
| NNDSVDA          | Nonnegative Double Singular Value Decomposition with zeros filled with the average          |
| NPMI             | Normalized Pointwise Mutual Information                                                     |

|                   |                                                         |
|-------------------|---------------------------------------------------------|
| OJK               | Otoritas Jasa Keuangan                                  |
| PDF               | Portable Document Format                                |
| PMI               | Pointwise Mutual Information                            |
| RAM               | Random Access Memory                                    |
| SDK               | Software Development Kit                                |
| SVD               | Singular Value Decomposition                            |
| TF                | Term Frequency                                          |
| TF-IDF            | Term Frequency-Inverse Document Frequency               |
| U <sub>mass</sub> | Coherence measure berbasis frekuensi kemunculan bersama |
| URL               | Uniform Resource Locator                                |
| UTF-8             | Unicode Transformation Format - 8-bit                   |
| V                 | Matriks dokumen-kata dalam NMF                          |
| W                 | Matriks dokumen-topik dalam NMF                         |
| XML               | eXtensible Markup Language                              |

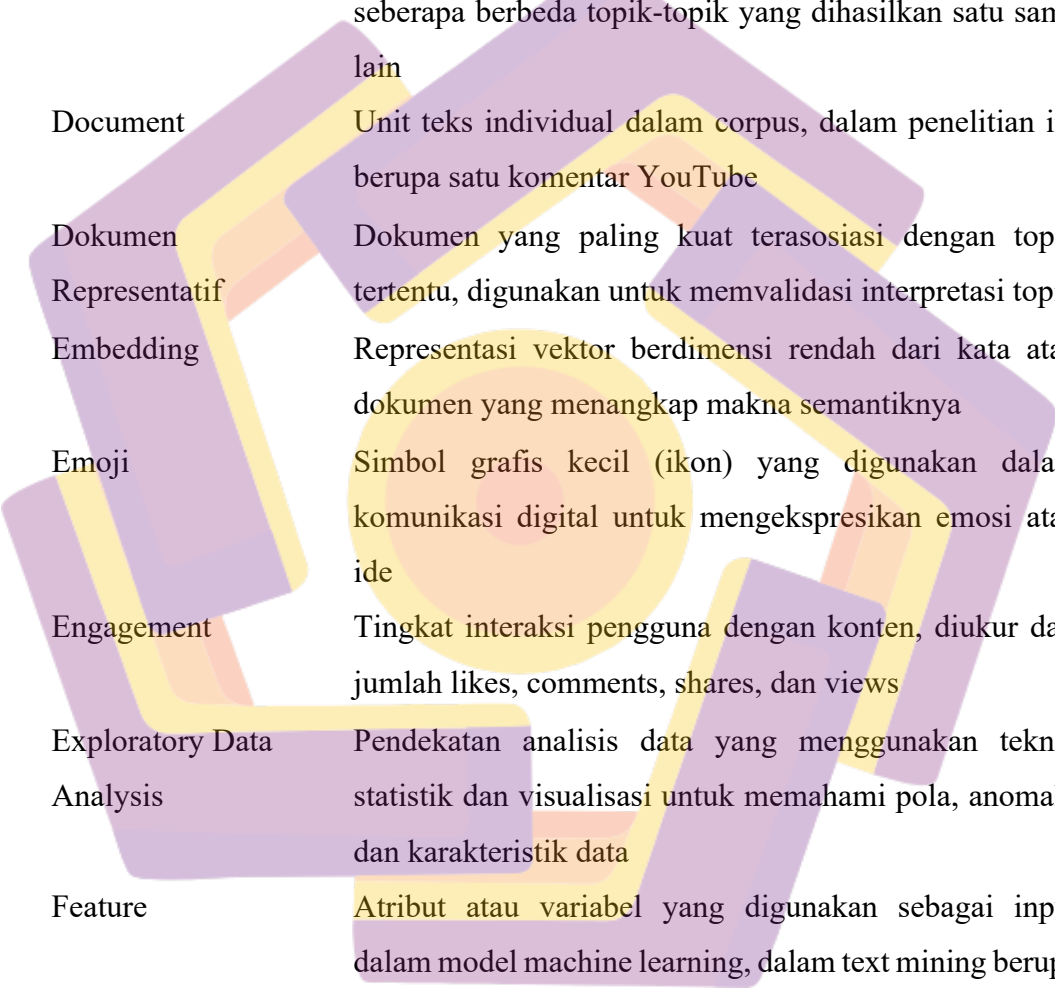
## DAFTAR ISTILAH



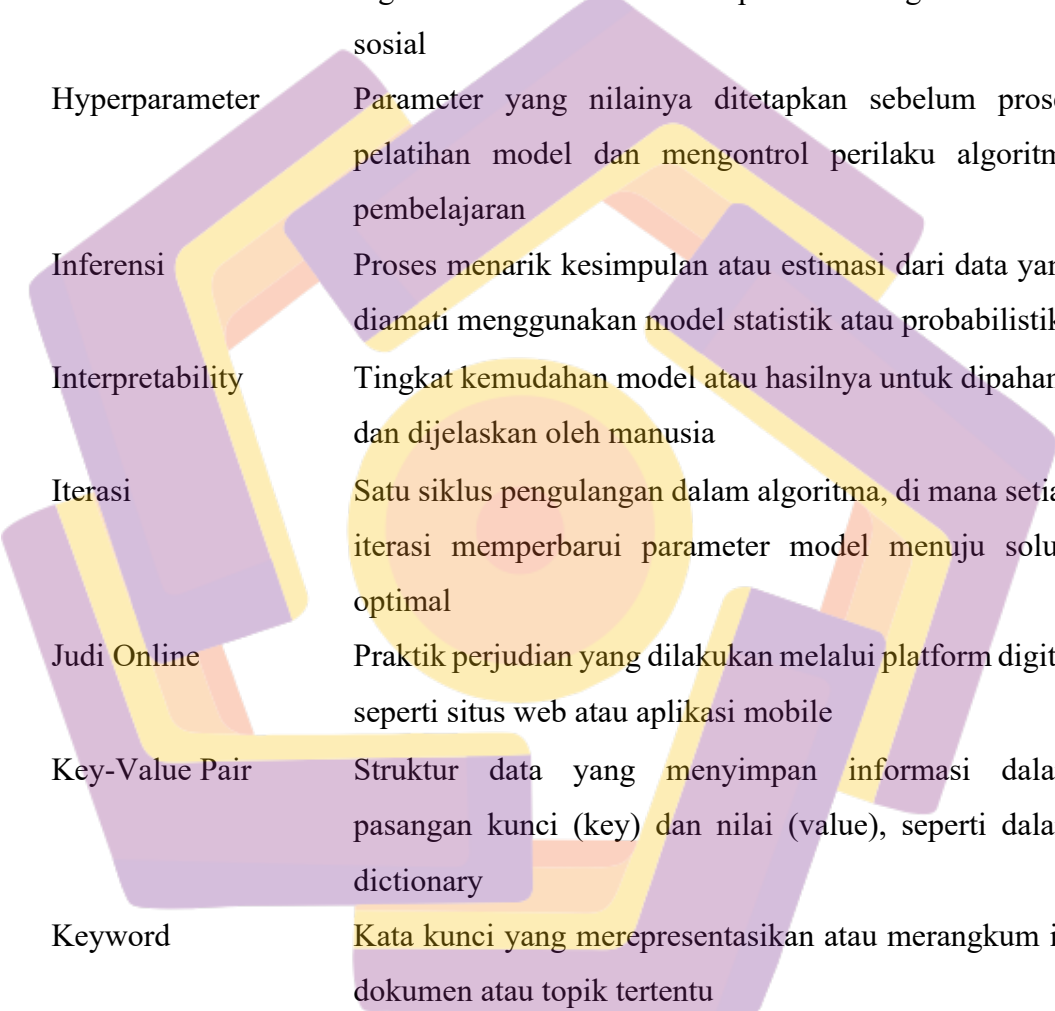
|                  |                                                                                                                                              |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| Ablation Study   | Eksperimen sistematis untuk mengukur kontribusi setiap komponen dalam sistem dengan menghapus komponen satu per satu                         |
| Algorithm        | Sekumpulan instruksi atau langkah-langkah terstruktur untuk menyelesaikan masalah atau melakukan komputasi                                   |
| API Key          | Kode unik yang digunakan untuk autentikasi saat mengakses layanan API, berfungsi sebagai identitas dan kontrol akses                         |
| Asymmetric Prior | Distribusi probabilitas awal yang memberikan bobot berbeda untuk setiap elemen, berbeda dengan symmetric prior yang memberikan bobot seragam |
| Baseline         | Model atau metode sederhana yang digunakan sebagai pembandingan untuk mengevaluasi performa model yang lebih kompleks                        |
| Bigram           | Pasangan dua kata yang muncul secara berurutan dalam teks, digunakan untuk menangkap frasa atau kolokasi bermakna                            |
| Bot              | Program komputer otomatis yang dirancang untuk melakukan tugas berulang, sering digunakan untuk posting komentar spam di media sosial        |
| Case Folding     | Proses konversi semua huruf dalam teks menjadi huruf kecil (lowercase) atau huruf besar (uppercase) untuk konsistensi                        |
| Channel          | Akun kreator konten di platform seperti YouTube yang memiliki koleksi video dan subscriber                                                   |
| Clustering       | Teknik machine learning untuk mengelompokkan data ke dalam kelompok (cluster) berdasarkan kemiripan karakteristik                            |



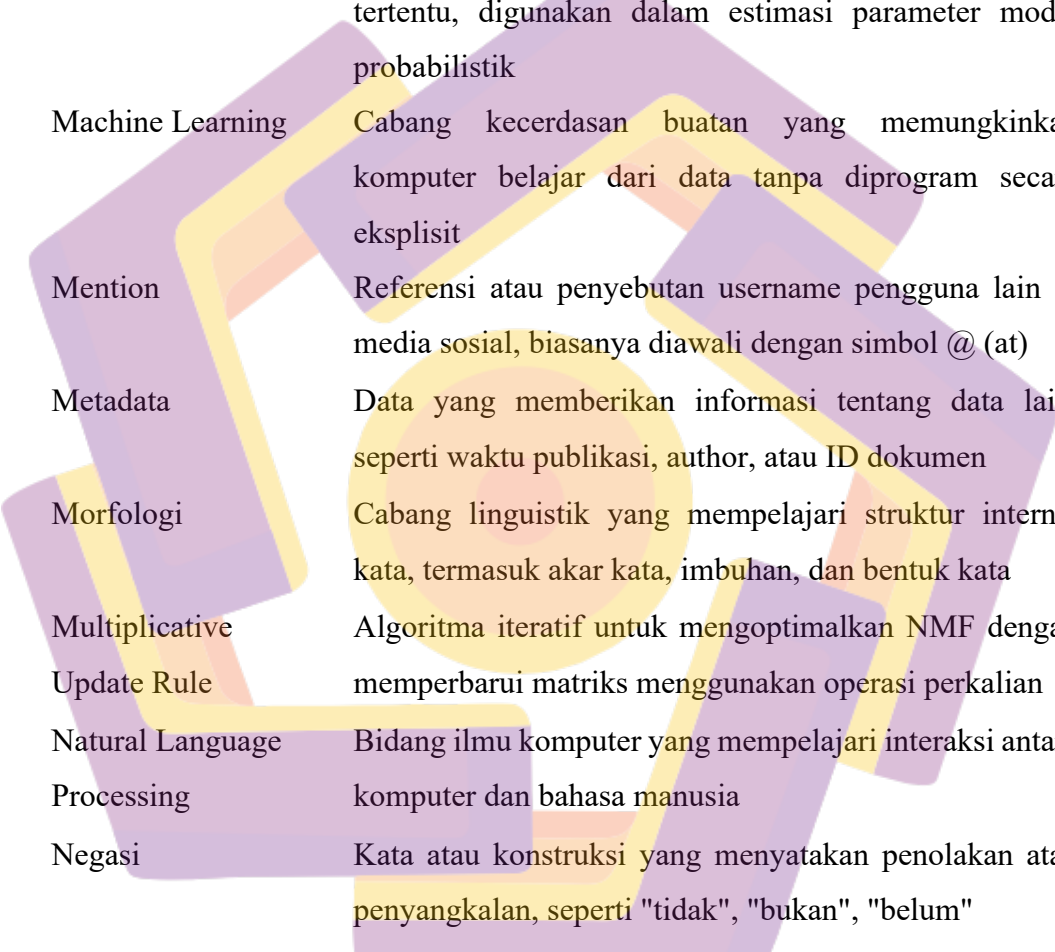
|                     |                                                                                                                                 |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Coherence Score     | Metrik untuk mengukur keterkaitan semantik antar kata dalam suatu topik, semakin tinggi nilainya semakin koheren topik tersebut |
| Colocation          | Pasangan atau kelompok kata yang sering muncul bersama dalam konteks tertentu dan memiliki makna khusus                         |
| Context Window      | Jendela teks di sekitar kata tertentu yang digunakan untuk menangkap konteks semantik kata tersebut                             |
| Convergence         | Kondisi ketika algoritma iteratif mencapai titik stabil dan tidak mengalami perubahan signifikan pada iterasi selanjutnya       |
| Corpus              | Kumpulan dokumen atau teks yang digunakan sebagai data untuk analisis bahasa atau pemodelan topik (jamak: corpora)              |
| Cosine Similarity   | Ukuran kemiripan antara dua vektor berdasarkan kosinus sudut di antara mereka, bernilai antara -1 hingga 1                      |
| Crawling            | Proses pengumpulan data dari website secara otomatis menggunakan program komputer (web crawler/spider)                          |
| Cross-reference     | Rujukan silang dalam dokumen yang menghubungkan satu bagian dengan bagian lain, sering digunakan untuk tabel dan gambar         |
| Dataset             | Kumpulan data terstruktur yang digunakan untuk analisis, pelatihan model, atau penelitian                                       |
| Dekomposisi Matriks | Teknik matematika untuk memecah matriks menjadi hasil perkalian dua atau lebih matriks yang lebih sederhana                     |
| Deterministik       | Algoritma atau proses yang menghasilkan output yang sama untuk input yang sama, tanpa elemen acak (randomness)                  |



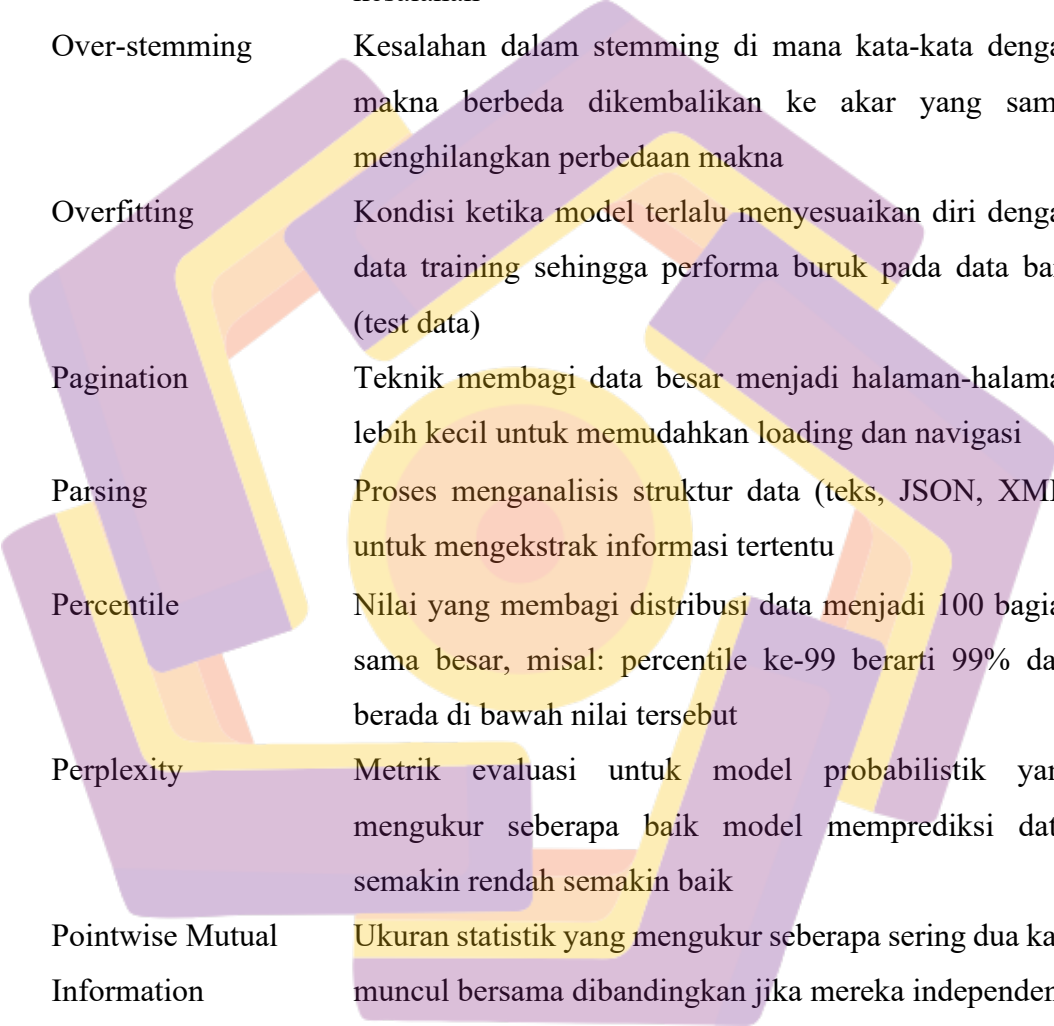
|                           |                                                                                                                                 |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Dictionary                | Struktur data yang memetakan kata-kata unik dalam corpus ke ID numerik, digunakan dalam representasi teks                       |
| Dirichlet Distribution    | Distribusi probabilitas kontinu multivariat yang sering digunakan sebagai prior dalam model Bayesian seperti LDA                |
| Diversity                 | Ukuran keberagaman antar topik, mengindikasikan seberapa berbeda topik-topik yang dihasilkan satu sama lain                     |
| Document                  | Unit teks individual dalam corpus, dalam penelitian ini berupa satu komentar YouTube                                            |
| Dokumen Representatif     | Dokumen yang paling kuat terasosiasi dengan topik tertentu, digunakan untuk memvalidasi interpretasi topik                      |
| Embedding                 | Representasi vektor berdimensi rendah dari kata atau dokumen yang menangkap makna semantiknya                                   |
| Emoji                     | Simbol grafis kecil (ikon) yang digunakan dalam komunikasi digital untuk mengekspresikan emosi atau ide                         |
| Engagement                | Tingkat interaksi pengguna dengan konten, diukur dari jumlah likes, comments, shares, dan views                                 |
| Exploratory Data Analysis | Pendekatan analisis data yang menggunakan teknik statistik dan visualisasi untuk memahami pola, anomali, dan karakteristik data |
| Feature                   | Atribut atau variabel yang digunakan sebagai input dalam model machine learning, dalam text mining berupa kata atau n-gram      |
| Frobenius Norm            | Ukuran besarnya matriks yang dihitung sebagai akar kuadrat dari jumlah kuadrat semua elemennya                                  |
| Generative Model          | Model probabilistik yang memodelkan bagaimana data dihasilkan dari distribusi probabilitas tertentu                             |



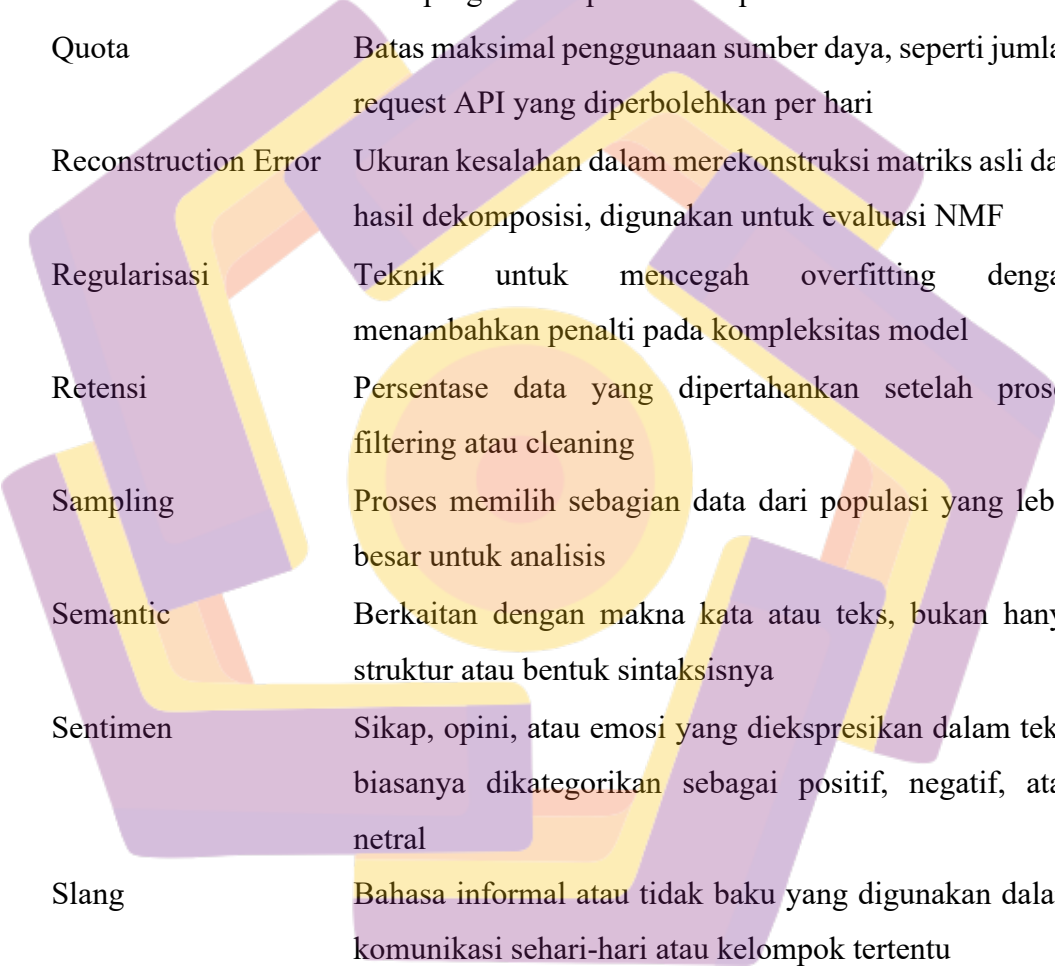
|                  |                                                                                                                       |
|------------------|-----------------------------------------------------------------------------------------------------------------------|
| Gibbs Sampling   | Algoritma inferensi untuk estimasi distribusi probabilitas dalam model Bayesian seperti LDA                           |
| Ground Truth     | Data yang dianggap benar atau akurat, digunakan sebagai referensi untuk evaluasi model supervised learning            |
| Hashtag          | Kata atau frasa yang diawali dengan tanda pagar (#) yang digunakan untuk menandai topik atau kategori di media sosial |
| Hyperparameter   | Parameter yang nilainya ditetapkan sebelum proses pelatihan model dan mengontrol perilaku algoritma pembelajaran      |
| Inferensi        | Proses menarik kesimpulan atau estimasi dari data yang diamati menggunakan model statistik atau probabilistik         |
| Interpretability | Tingkat kemudahan model atau hasilnya untuk dipahami dan dijelaskan oleh manusia                                      |
| Iterasi          | Satu siklus pengulangan dalam algoritma, di mana setiap iterasi memperbarui parameter model menuju solusi optimal     |
| Judi Online      | Praktik perjudian yang dilakukan melalui platform digital seperti situs web atau aplikasi mobile                      |
| Key-Value Pair   | Struktur data yang menyimpan informasi dalam pasangan kunci (key) dan nilai (value), seperti dalam dictionary         |
| Keyword          | Kata kunci yang merepresentasikan atau merangkum isi dokumen atau topik tertentu                                      |
| Kolokasi         | Kata-kata yang cenderung muncul bersama secara berulang dan membentuk makna khusus (synonym: colocation)              |
| Latent           | Tersembunyi atau tidak langsung teramati, dalam topic modeling merujuk pada topik yang diekstrak dari pola kata       |



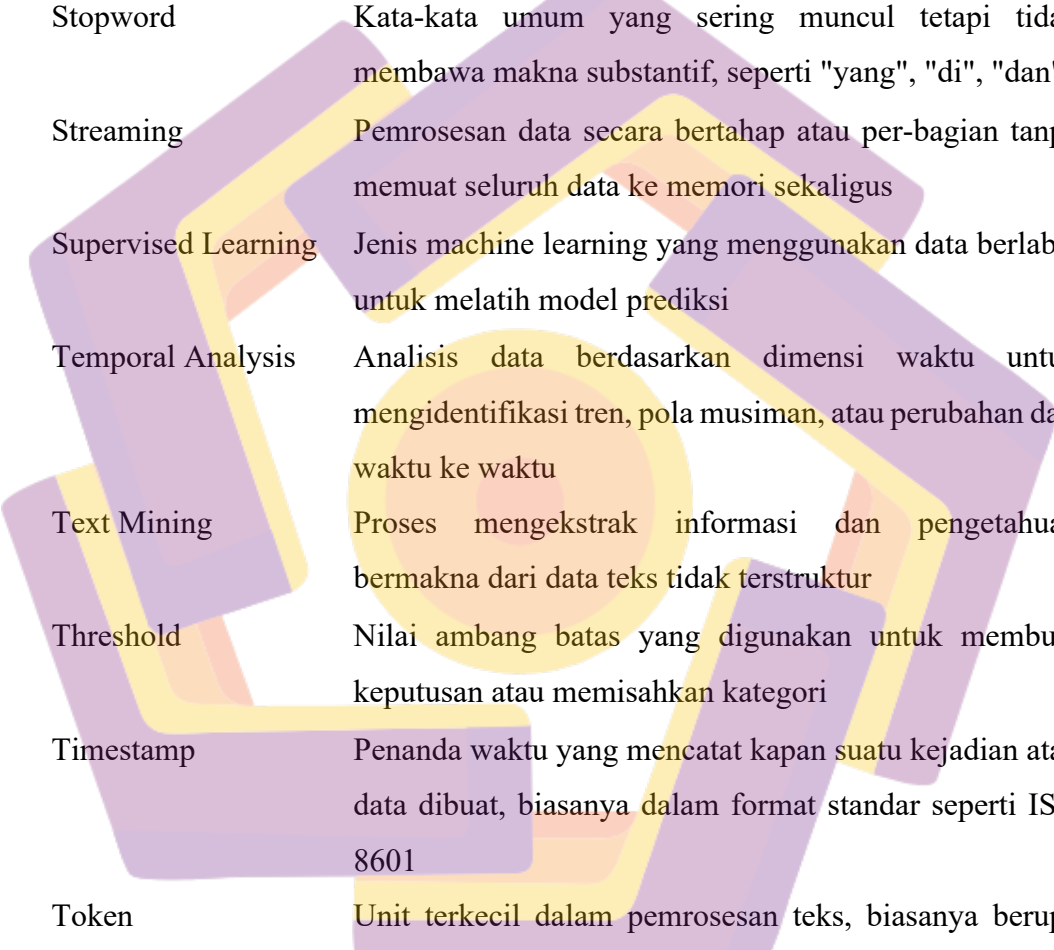
|                             |                                                                                                                                 |
|-----------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Learning Curve              | Grafik yang menunjukkan perubahan performa model terhadap jumlah data training atau iterasi pembelajaran                        |
| Lemmatization               | Proses mengubah kata ke bentuk dasar (lemma) berdasarkan analisis morfologi dan kamus, lebih akurat dari stemming               |
| Likelihood                  | Probabilitas mengamati data yang diberikan dalam model tertentu, digunakan dalam estimasi parameter model probabilistik         |
| Machine Learning            | Cabang kecerdasan buatan yang memungkinkan komputer belajar dari data tanpa diprogram secara eksplisit                          |
| Mention                     | Referensi atau penyebutan username pengguna lain di media sosial, biasanya diawali dengan simbol @ (at)                         |
| Metadata                    | Data yang memberikan informasi tentang data lain, seperti waktu publikasi, author, atau ID dokumen                              |
| Morfologi                   | Cabang linguistik yang mempelajari struktur internal kata, termasuk akar kata, imbuhan, dan bentuk kata                         |
| Multiplicative Update Rule  | Algoritma iteratif untuk mengoptimalkan NMF dengan memperbarui matriks menggunakan operasi perkalian                            |
| Natural Language Processing | Bidang ilmu komputer yang mempelajari interaksi antara komputer dan bahasa manusia                                              |
| Negasi                      | Kata atau konstruksi yang menyatakan penolakan atau penyangkalan, seperti "tidak", "bukan", "belum"                             |
| N-gram                      | Sekuens berurutan dari n item (biasanya kata) dari teks, misal: unigram (1 kata), bigram (2 kata), trigram (3 kata)             |
| Noise                       | Data yang tidak relevan, tidak akurat, atau mengganggu analisis, seperti spam atau kesalahan input                              |
| Normalisasi                 | Proses mengubah data ke bentuk standar atau rentang tertentu, dalam text mining berarti mengubah kata tidak baku ke bentuk baku |



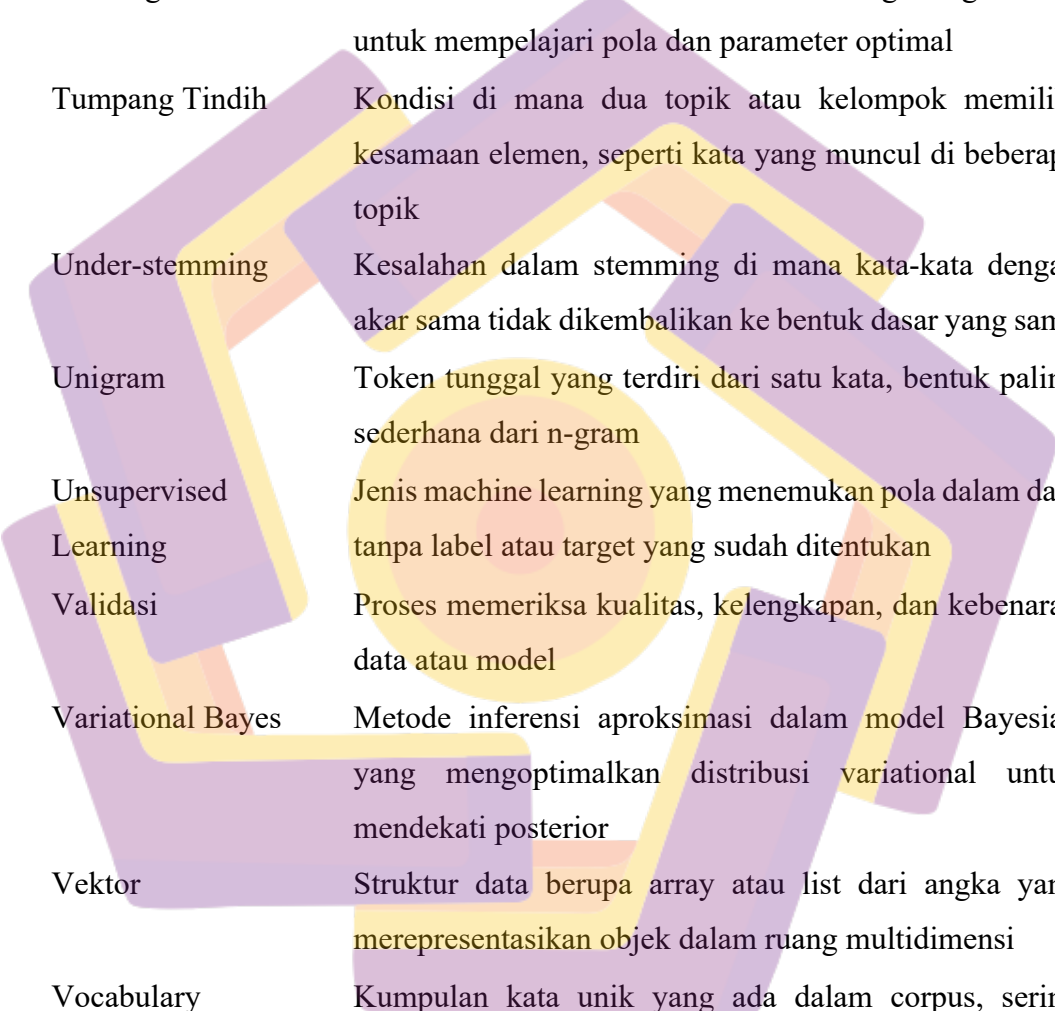
|                              |                                                                                                                                           |
|------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Optimasi                     | Proses mencari solusi terbaik (optimal) untuk masalah tertentu dengan meminimalkan atau memaksimalkan fungsi objektif                     |
| Outlier                      | Nilai data yang sangat berbeda atau jauh dari sebagian besar data lainnya, bisa merupakan anomali atau kesalahan                          |
| Over-stemming                | Kesalahan dalam stemming di mana kata-kata dengan makna berbeda dikembalikan ke akar yang sama, menghilangkan perbedaan makna             |
| Overfitting                  | Kondisi ketika model terlalu menyesuaikan diri dengan data training sehingga performa buruk pada data baru (test data)                    |
| Pagination                   | Teknik membagi data besar menjadi halaman-halaman lebih kecil untuk memudahkan loading dan navigasi                                       |
| Parsing                      | Proses menganalisis struktur data (teks, JSON, XML) untuk mengekstrak informasi tertentu                                                  |
| Percentile                   | Nilai yang membagi distribusi data menjadi 100 bagian sama besar, misal: percentile ke-99 berarti 99% data berada di bawah nilai tersebut |
| Perplexity                   | Metrik evaluasi untuk model probabilistik yang mengukur seberapa baik model memprediksi data, semakin rendah semakin baik                 |
| Pointwise Mutual Information | Ukuran statistik yang mengukur seberapa sering dua kata muncul bersama dibandingkan jika mereka independen                                |
| Preprocessing                | Tahap awal dalam analisis data yang meliputi pembersihan, transformasi, dan standarisasi data mentah                                      |
| Prevalensi                   | Proporsi atau persentase kemunculan suatu topik dalam keseluruhan corpus                                                                  |



|                      |                                                                                                                      |
|----------------------|----------------------------------------------------------------------------------------------------------------------|
| Prior Distribution   | Distribusi probabilitas awal yang mencerminkan belief sebelum mengamati data, digunakan dalam inferensi Bayesian     |
| Probabilistik        | Berdasarkan probabilitas atau kemungkinan, tidak deterministik                                                       |
| Pseudocode           | Deskripsi algoritma dalam bahasa informal yang mirip kode program tetapi mudah dipahami manusia                      |
| Quota                | Batas maksimal penggunaan sumber daya, seperti jumlah request API yang diperbolehkan per hari                        |
| Reconstruction Error | Ukuran kesalahan dalam merekonstruksi matriks asli dari hasil dekomposisi, digunakan untuk evaluasi NMF              |
| Regularisasi         | Teknik untuk mencegah overfitting dengan menambahkan penalti pada kompleksitas model                                 |
| Retensi              | Persentase data yang dipertahankan setelah proses filtering atau cleaning                                            |
| Sampling             | Proses memilih sebagian data dari populasi yang lebih besar untuk analisis                                           |
| Semantic             | Berkaitan dengan makna kata atau teks, bukan hanya struktur atau bentuk sintaksisnya                                 |
| Sentimen             | Sikap, opini, atau emosi yang diekspresikan dalam teks, biasanya dikategorikan sebagai positif, negatif, atau netral |
| Slang                | Bahasa informal atau tidak baku yang digunakan dalam komunikasi sehari-hari atau kelompok tertentu                   |
| Spam                 | Konten yang tidak diinginkan, sering berupa iklan atau pesan berulang yang mengganggu                                |
| Sparse Matrix        | Matriks yang sebagian besar elemennya bernilai nol, disimpan secara efisien dengan hanya menyimpan elemen non-zero   |



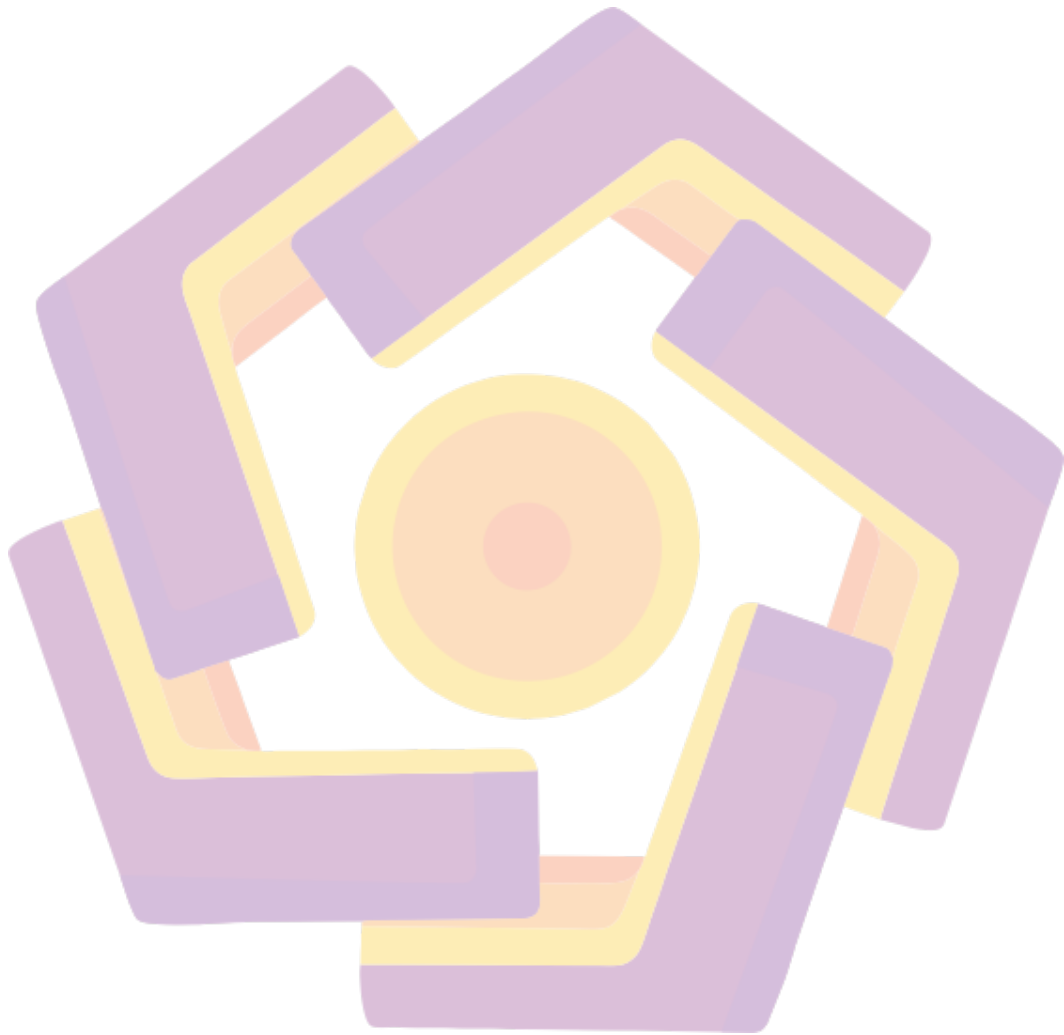
|                     |                                                                                                                       |
|---------------------|-----------------------------------------------------------------------------------------------------------------------|
| Sparsity            | Tingkat kejarangan dalam data, dihitung sebagai proporsi elemen nol dalam matriks atau vektor                         |
| Stemming            | Proses mengubah kata berimbuhan menjadi bentuk dasar (root/stem) dengan menghapus imbuhan                             |
| Stokastik           | Memiliki elemen acak atau probabilitas, hasil dapat berbeda meskipun input sama                                       |
| Stopword            | Kata-kata umum yang sering muncul tetapi tidak membawa makna substantif, seperti "yang", "di", "dan"                  |
| Streaming           | Pemrosesan data secara bertahap atau per-bagian tanpa memuat seluruh data ke memori sekaligus                         |
| Supervised Learning | Jenis machine learning yang menggunakan data berlabel untuk melatih model prediksi                                    |
| Temporal Analysis   | Analisis data berdasarkan dimensi waktu untuk mengidentifikasi tren, pola musiman, atau perubahan dari waktu ke waktu |
| Text Mining         | Proses mengekstrak informasi dan pengetahuan bermakna dari data teks tidak terstruktur                                |
| Threshold           | Nilai ambang batas yang digunakan untuk membuat keputusan atau memisahkan kategori                                    |
| Timestamp           | Penanda waktu yang mencatat kapan suatu kejadian atau data dibuat, biasanya dalam format standar seperti ISO 8601     |
| Token               | Unit terkecil dalam pemrosesan teks, biasanya berupa kata individual setelah proses tokenisasi                        |
| Tokenisasi          | Proses memecah teks menjadi unit-unit lebih kecil (token), biasanya kata atau karakter                                |
| Topic               | Tema atau subjek yang direpresentasikan sebagai distribusi probabilitas atas kata-kata dalam topic modeling           |



|                       |                                                                                                                        |
|-----------------------|------------------------------------------------------------------------------------------------------------------------|
| Topic Modeling        | Teknik unsupervised machine learning untuk menemukan struktur tematik tersembunyi dalam kumpulan dokumen               |
| Trade-off             | Situasi di mana peningkatan satu aspek menyebabkan penurunan aspek lain, memerlukan kompromi                           |
| Training              | Proses melatih model machine learning dengan data untuk mempelajari pola dan parameter optimal                         |
| Tumpang Tindih        | Kondisi di mana dua topik atau kelompok memiliki kesamaan elemen, seperti kata yang muncul di beberapa topik           |
| Under-stemming        | Kesalahan dalam stemming di mana kata-kata dengan akar sama tidak dikembalikan ke bentuk dasar yang sama               |
| Unigram               | Token tunggal yang terdiri dari satu kata, bentuk paling sederhana dari n-gram                                         |
| Unsupervised Learning | Jenis machine learning yang menemukan pola dalam data tanpa label atau target yang sudah ditentukan                    |
| Validasi              | Proses memeriksa kualitas, kelengkapan, dan kebenaran data atau model                                                  |
| Variational Bayes     | Metode inferensi aproksimasi dalam model Bayesian yang mengoptimalkan distribusi variational untuk mendekati posterior |
| Vektor                | Struktur data berupa array atau list dari angka yang merepresentasikan objek dalam ruang multidimensi                  |
| Vocabulary            | Kumpulan kata unik yang ada dalam corpus, sering disebut juga kosakata                                                 |
| Weight                | Bobot atau nilai numerik yang menunjukkan kepentingan atau kontribusi suatu elemen dalam model                         |
| Word Cloud            | Visualisasi data teks di mana ukuran kata proporsional terhadap frekuensi atau kepentingannya                          |

Word Embedding

Representasi vektor kontinu berdimensi rendah dari kata yang menangkap hubungan semantik dan sintaktik



## INTISARI

Fenomena judi online di Indonesia menunjukkan peningkatan yang signifikan dan banyak bermunculan promosi terselubung di ruang publik digital, sehingga menimbulkan risiko sosial-ekonomi dan tantangan bagi upaya penegakan kebijakan. Penelitian ini bertujuan memetakan isu terkait judi online pada komentar publik YouTube dengan membandingkan dua pendekatan pemodelan topik: Latent Dirichlet Allocation (LDA) dan Non-Negative Matrix Factorization (NMF). Data dikumpulkan melalui YouTube Data API v3 pada rentang 30 Agustus 2023–7 Desember 2025 dan setelah validasi diperoleh 6.205 komentar top-level. Proses analisis meliputi pra-pemrosesan teks (case folding, pembersihan URL/emoji, normalisasi slang, tokenisasi, stopword removal, stemming, dan pembentukan bigram), representasi BoW untuk LDA serta TF-IDF untuk NMF, pelatihan model dengan variasi jumlah topik, dan evaluasi menggunakan metrik koherensi serta inspeksi manual untuk interpretasi topik dan visualisasi temporal. Hasil menunjukkan beberapa topik dominan seperti modus promosi (tautan/afiliasi), istilah permainan (slot, togel), mekanisme deposit/withdraw, peran influencer, dan kritik publik terhadap dampak ekonomi; NMF cenderung menghasilkan topik yang lebih mudah diinterpretasi sementara LDA efektif menangkap campuran topik dalam komentar. Kontribusi penelitian meliputi pemetaan tematik isu judi online, dataset terstruktur untuk penelitian lanjutan, serta rekomendasi bagi pembuat kebijakan dan moderator platform untuk deteksi dan mitigasi konten. Rekomendasi lanjutan mencakup penerapan teknik embedding berbasis transformer dan perluasan sumber data lintas platform.

**Kata kunci:** Judi Online, Pemodelan Topik, LDA, NMF, YouTube

## **ABSTRACT**

*The rapid growth of online gambling in Indonesia has intensified the spread of hidden promotional content across digital public spaces, creating socio-economic risks and challenges for policy enforcement. This study aims to map issues related to online gambling through public YouTube comments by comparing two topic modeling approaches: Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Data were collected using the YouTube Data API v3 within the period of August 30, 2023 to December 7, 2025, resulting in 6,205 validated top-level comments. The analysis process included text preprocessing (case folding, removal of URLs and emojis, slang normalization, tokenization, stopword removal, stemming, and bigram formation), text representation using Bag-of-Words for LDA and TF-IDF for NMF, model training with variations in the number of topics, and evaluation using coherence metrics combined with manual interpretation and temporal visualization. The results reveal dominant themes such as promotional strategies (links and affiliates), gambling-related terms (slot and togel), deposit and withdrawal mechanisms, the role of influencers, and public criticism regarding economic impacts. NMF tends to produce more interpretable topics, while LDA effectively captures topic mixtures within comments. This research contributes a thematic mapping of online gambling discourse, a structured dataset for future studies, and practical recommendations for policymakers and platform moderators in detecting and mitigating harmful digital content.*

**Keyword:** Online Gambling, Topic Modeling, LDA, NMF, YouTube