BAB I

PENDAHULUAN

1.1 Latar Belakang

Fenomena judi *online* menunjukkan perkembangan yang signifikan di Indonesia, sejalan dengan meningkatnya akses internet dan kemajuan teknologi digital. Aktivitas ini menjadi isu yang meresahkan karena menimbulkan berbagai dampak negatif yang serius. Judi *online* dinilai berbahaya karena dapat memengaruhi berbagai aspek kehidupan masyarakat, termasuk sosial, ekonomi, dan psikologis (Laras et al., 2024).

Menurut Pusat Pelaporan dan Analisis Keuangan (PPATK) pada tahun 2025, perputaran dana judi *online* yang tercatat telah mencapai 1.200 Trilliun, data ini menunjukkan peningkatan dibandingkan tahun sebelumnya sebesar 981 trilliun (Akbar, 2025). Selain itu menurut Menkopolkan Budi Gunawan pada tahun 2024 mengatakan bahwa jumlah penduduk Indonesia yang bermain judi *online* telah mencapai 8,8 juta orang, dimana 80 ribu diantaranya adalah anak-anak. Fakta tersebut menunjukkan bahwa penyebaran judi *online* sudah sangat marak di tengah masyarakat. Faktor-faktor seperti kemudahan akses dan strategi promosi yang agresif telah berkontribusi secara signifikan terhadap meningkatnya tingkat kecanduan terhadap judi *online* (Laras et al., 2024).

Promosi perjudian *online* biasanya memanfaatkan media sosial, layanan *streaming*, dan blog untuk menjangkau pengguna internet dengan mengiklankan situs permainan mereka. Pemilik situs perjudian *online* kerap menawarkan biaya promosi yang lebih besar kepada pemilik akun media sosial dibandingkan dengan biaya yang biasanya dikeluarkan untuk iklan konvensional. Namun, semenjak ditetapkan Keputusan Presiden Nomor 21 Tahun 2024 tentang Satuan Tugas Pemberantasan Perjudian Daring (Peraturan Pemerintah RI, 2024), promosi perjudian *online* melalui *endorsement* akun media sosial terkenal mulai mengalami penurunan seiring meningkatnya kesadaran pemerintah dan masyarakat dalam memerangi penyebaran perjudian daring (Majid & Maskur, 2023). Pemilik akun media sosial juga semakin enggan menerima tawaran tersebut, meskipun bernilai besar, karena adanya tekanan sosial dan potensi sanksi hukum yang berat apabila diketahui terlibat dalam promosi aktivitas ilegal tersebut (Irwandi, 2025).

Hal tersebut belum cukup untuk mengurangi penyebaran promosi judi *online*. Para pelaku menggunakan berbagai cara kreatif untuk memperkenalkan situsnya, dimana salah satunya melalui cara spam di berbagai komentar video youtube. Pola perilaku promosi yang dilakukan oleh pelaku cenderung mirip di setiap video youtube. Seringkali akun yang digunakan untuk melakukan promosi adalah akun baru. Selain itu, pola semantik yang digunakan juga mirip karena berisi ajakan tersirat untuk bermain di situs judi *online*. Dengan miripnya pola tersebut ada kecenderungan bahwa para pelaku promosi tersebut menggunakan sistem otomatis atau *bot* karena memiliki ciri yang sama, jumlahnya yang sangat banyak, dan waktu komentarnya yang saling berdekatan dalam setiap video.

Menurut Bening (2025), metode yang dilakukan oleh pelaku judi *online* tersebut sangat meresahkan pemilik akun youtube dan juga para penontonnya karena kolom komentar yang seharusnya menjadi media berdiskusi telah dipenuhi oleh promosi judi *online*. Para pemilik akun merasa kesulitan untuk menghapus komentar tersebut karena harus membaca satu per satu komentar dan menghapus secara manual promosi judi *online* tersebut. Penonton video juga sulit berkontribusi untuk melakukan *report* terhadap promosi judi *online* tersebut karena belum ada sistem resmi dari youtube untuk mendeteksi promosi judi *online* tersebut.

Salah satu pendekatan yang dapat dilakukan untuk mengatasi penyebaran komentar promosi judi *online* adalah dengan membangun sistem deteksi spam yang tidak hanya mempertimbangkan aspek semantik dari komentar, tetapi juga memperhitungkan pola perilaku pengguna yang menyebarkan komentar tersebut. Pola perilaku antara lain seperti, umur akun pengguna, jumlah *likes* yang diterima, jumlah karakter tidak lazim yang digunakan dalam komentar, serta interval waktu antar komentar yang diduga promosi judi *online*. Sistem ini juga perlu dirancang menggunakan arsitektur *deep learning* karena komentar promosi sering kali disampaikan secara tersirat, tanpa menggunakan kata kunci yang eksplisit. Kondisi ini menyulitkan sistem deteksi berbasis *rule-based* seperti *pattern matching* menggunakan *regular expression*, yang sangat bergantung pada keberadaan kata atau frasa tertentu dalam teks sebagaimana yang dilakukan oleh Cahyo et al. (2025).

Menurut penelitian yang dilakukan oleh Alhassun & Rassam (2022), evaluasi terhadap berbagai arsitektur *deep learning* menunjukkan bahwa model yang menggabungkan fitur semantik dan fitur perilaku yang diperoleh dari metadata pengguna mampu memberikan akurasi yang lebih tinggi dibandingkan model yang hanya mengandalkan fitur semantik. Temuan ini menunjukkan bahwa pendekatan *hybrid* lebih efektif dalam menangkap kompleksitas pola penyebaran komentar promosi yang bersifat implisit.

Selain itu, Zouak et al. (2025) pada penelitian deteksi spam emailnya menyatakan bahwa pendekatan konvensional pada umumnya hanya bergantung pada fitur tekstual dan sering mengabaikan kesamaan struktural laten antar email, seperti pola format yang berulang atau kedekatan semantik. Penelitian tersebut mengatasi hal ini dengan mengombinasikan *Bidirectional Encoder Representations from Transformers* (BERT) untuk pemahaman semantik dan *Graph Sample and Aggregate* (GraphSAGE) untuk menangkap relasi antar-email. Meskipun demikian, model ini masih memiliki beberapa keterbatasan karena hanya bergantung pada konten tekstual dan belum memanfaatkan metadata struktural dari pengguna.

Oleh karena itu, penelitian ini akan dilakukan untuk mengembangkan sistem deteksi promosi judi *online* dengan mempertimbangkan pola komentar serta perilaku pengguna. Analisis ini bertujuan untuk mengeksplorasi celah yang belum dibahas dalam penelitian sebelumnya. Arsitektur yang akan digunakan adalah kombinasi IndoBERTweet-GraphSAGE. IndoBERTweet dipilih karena telah di *fine-tune* menggunakan beragam data tweet dari platform X sehingga lebih baik dalam memahami teks informal berbahasa Indonesia. Hal ini sejalan dengan temuan Indriani et al. (2023) yang menyatakan bahwa IndoBERTweet memiliki performa yang unggul dalam

menangani teks berbahasa Indonesia yang bersifat informal dan tidak baku, seperti yang umum ditemukan di media sosial.

Sistem ini nantinya akan diimplementasikan dalam bentuk aplikasi web yang ditujukan untuk membantu pemilik akun YouTube dan moderator dalam menghapus komentar promosi judi *online* secara efisien, tanpa perlu melakukan pemilahan secara manual satu per satu. Selain itu, sistem ini juga dapat digunakan oleh penonton video YouTube untuk melaporkan komentar promosi judi *online*, sehingga turut membantu pemilik channel, moderator, dan sistem YouTube dalam proses deteksi. Harapannya, sistem ini dapat berfungsi sebagai media kolaboratif dalam mencegah penyebaran konten promosi judi *online*. Untuk memastikan bahwa aplikasi web ini berjalan sesuai dengan fungsionalitas yang telah dirancang, maka dilakukan metode pengujian *blackbox* yang berfokus pada pengujian *input* dan *output* tanpa melihat struktur internal dari sistem. Pengujian ini menggunakan salah satu teknik *blackbox*, yaitu *use case testing* yang bertujuan untuk mengevaluasi apakah seluruh fitur berjalan sebagaimana mestinya dari perspektif pengguna.

1.2 Landasan Teori

1.2.1 *Natural Languange Processing* (NLP)

Natural Language Processing (NLP) merupakan subbidang dari ilmu komputer dan kecerdasan buatan yang berfokus pada pemrosesan bahasa manusia oleh komputer (Stryker & Holdsworth, 2024). Definisi tersebut menegaskan bahwa NLP memanfaatkan teknik *machine learning* untuk memungkinkan sistem komputasi memahami makna bahasa alami serta berinteraksi dengan manusia secara lebih natural. Dalam konteks penelitian, NLP berperan penting dalam mentransformasikan teks tidak terstruktur menjadi representasi numerik yang dapat dianalisis menggunakan algoritma pembelajaran mesin maupun pembelajaran mendalam. Proses NLP umumnya mencakup beberapa tahapan utama, seperti prapemrosesan teks, ekstraksi fitur, serta pembentukan representasi numerik, sehingga informasi linguistik dalam teks dapat diolah secara komputasional untuk mendukung berbagai tugas analisis.

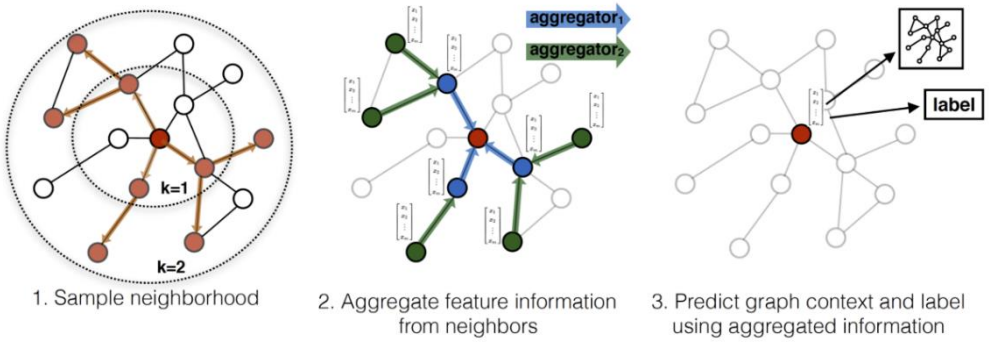
1.2.2 *Graph Neural Network* (GNN)

Graph Neural Network (GNN) merupakan pendekatan *deep learning* yang dirancang untuk memodelkan dan menganalisis data yang direpresentasikan dalam bentuk graf. Dalam konteks ini, graf merupakan struktur data yang digunakan untuk merepresentasikan sekumpulan objek yang dinyatakan sebagai *node* serta hubungan antar objek tersebut yang direpresentasikan dalam bentuk *edge* (Zhou et al., 2020). Berbeda dengan metode deep learning konvensional yang bekerja pada data berstruktur tetap seperti citra atau teks berurutan, GNN mampu menangkap ketergantungan relasional dan struktural antar entitas dalam graf. Mekanisme utama GNN adalah message passing, yaitu proses di mana setiap *node* secara iteratif mengagregasi informasi dari *node-node* tetangganya untuk memperbarui representasi fitur yang dimilikinya. Proses ini memungkinkan GNN untuk mempelajari representasi *node* yang

tidak hanya bergantung pada atribut lokal, tetapi juga pada konteks struktural dan hubungan antar *node* di dalam graf. Dengan kemampuan tersebut, GNN banyak digunakan pada berbagai tugas yang memiliki hubungan kompleks seperti jejaring sosial, sistem rekomendasi, dan analisis entitas berbasis relasi.

1.2.3 GraphSAGE

GraphSAGE merupakan salah satu varian model *Graph Neural Network* (GNN) yang dirancang untuk membentuk representasi atau *embedding node* yang lebih kontekstual pada graf. Berbeda dengan *Graph Convolutional Network* (GCN) yang bersifat transductive dan mengharuskan seluruh struktur graf diketahui selama proses pelatihan, GraphSAGE dikembangkan sebagai perluasan dari GCN untuk mengatasi keterbatasan tersebut. Hamilton et al. memperluas konsep konvolusi pada GCN ke dalam kerangka pembelajaran induktif, sehingga model tidak lagi bergantung pada identitas *node* tertentu maupun struktur graf secara keseluruhan. Dengan mengagregasi sampel fitur tetangga, GraphSAGE mampu menghasilkan representasi *node* berdasarkan fitur dan struktur lokalnya, sehingga dapat melakukan inferensi pada *node* baru yang tidak pernah muncul selama tahap pelatihan. Prinsip utama GraphSAGE adalah melakukan proses *sampling* terhadap sejumlah tetangga dari setiap *node*, kemudian menggabungkan (*aggregate*) informasi dari tetangga tersebut menggunakan fungsi agregasi seperti *mean*. Cara kerja GraphSAGE dapat diilustrasikan seperti pada Gambar 1.



Gambar 1. Cara kerja GraphSAGE (Hamilton et al., 2018)

Secara matematis, proses pembaruan representasi *node* v pada lapisan ke-k dalam GraphSAGE dinyatakan dengan persamaan (1)

$$h_v^{(k)} = \sigma(W^{(k)} \cdot [\, h_v^{(k-1)} \,\|\, \text{AGGREGATE}(\{\, h_u^{(k-1)}, \forall u \in \mathcal{N}(v)\}) \,]) \qquad (1)$$

dimana $h_v^{(k)}$ merepresentasikan *embedding* atau representasi *node* v pada lapisan ke-k, yang diperbarui secara iteratif dengan memanfaatkan informasi dari *node* itu sendiri dan tetangganya. Simbol $N(v)$ menyatakan himpunan tetangga dari *node* v, yaitu *node-node* yang memiliki hubungan langsung dengan v dalam graf. Selanjutnya, $h_u^{(k-1)}$ melambangkan representasi *node* tetangga u pada lapisan sebelumnya (ke-$k-1$), yang

digunakan sebagai sumber informasi kontekstual dalam proses agregasi. Matriks $W^{(k)}$ merupakan parameter bobot yang dapat dilatih pada lapisan ke-k, berfungsi untuk melakukan transformasi linear terhadap hasil gabungan antara representasi *node* vdan hasil agregasi tetangganya. Fungsi σ menunjukkan fungsi aktivasi non-linear, seperti ReLU, yang digunakan untuk menghasilkan non-linearitas pada pembaruan representasi. Operasi konkatenasi (‖) digunakan untuk menggabungkan representasi *node* $h_v^{(k-1)}$ dengan hasil fungsi agregasi tetangga sebelum dilakukan transformasi linear oleh $W^{(k)}$. Adapun fungsi AGGREGATE merupakan mekanisme penggabungan informasi dari tetangga-tetangga *node*. Adapun salah satu *aggregator* dalam GraphSage adalah seperti berikut.

Mean Aggregator, yang menghitung rata-rata representasi tetangga dinyatakan dalam notasi (2)

$$h_{\mathcal{N}(v)}^{(k)} = \frac{1}{|\,\mathcal{N}(v)\,|} \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)} \tag{2}$$

Dimana proses agregasi tetangga pada lapisan ke-k dalam model GraphSAGE dengan menggunakan pendekatan *mean aggregator*. Pada persamaan tersebut, $N(v)$ merepresentasikan himpunan tetangga dari *node* v, sedangkan $|\,N(v)\,|$ menyatakan jumlah tetangga yang dimiliki *node* tersebut. Setiap tetangga $u \in N(v)$ memiliki vektor representasi $h_u^{(k-1)}$ pada lapisan sebelumnya $(k-1)$, dan seluruh representasi tetangga ini dirata-ratakan untuk menghasilkan vektor agregasi $h_{N(v)}^{(k)}$. Nilai rata-rata ini menggambarkan informasi kolektif dari lingkungan lokal *node* v pada graf.

Menurut Hamilton et al., *mean aggregator* merupakan varian GraphSAGE yang paling efisien secara komputasi dibandingkan aggregator lainnya (pooling dan LSTM), baik pada tahap pelatihan (*training*) maupun inferensi. Mekanisme agregasi mean hanya melibatkan operasi perataan sederhana terhadap *embedding node* tetangga. Oleh karena itu, *mean aggregator* memiliki kompleksitas waktu yang lebih rendah serta skalabilitas yang lebih baik dibandingkan varian lainnya. Karakteristik ini menjadikan mean aggregator lebih sesuai untuk penerapan pada *graph* berskala besar dan data yang bersifat dinamis, di mana efisiensi komputasi dan kemampuan melakukan inferensi cepat pada *node* baru menjadi faktor yang sangat penting.

Hasil agregasi ini kemudian digabungkan dengan representasi *node* itu sendiri $h_v^{(k-1)}$ untuk menghasilkan *embedding* baru $h_v^{(k)}$ yang lebih informatif. Pada tahap ini, GraphSAGE hanya bertujuan mempelajari pola hubungan setiap *node* terhadap tetangganya melalui pendekatan *sample and aggregate*, dengan memanfaatkan fitur *node* dan struktur graf, belum melibatkan informasi label kelas untuk melakukan klasifikasi. Dengan demikian, model belum mengetahui apakah suatu *node* termasuk ke dalam kelas tertentu. Pemahaman model terhadap kecenderungan suatu *node* untuk dikategorikan ke dalam kelas tertentu baru terbentuk pada tahap pelatihan, yaitu ketika *embedding node* dari GraphSAGE diproyeksikan ke *fully connected layer*. Pada tahap ini, bobot $W^{(k)}$ akan disesuaikan secara bertahap selama proses pelatihan agar model

mampu mengkategorikan suatu *node* ke dalam kelas tertentu secara andal berdasarkan pola hubungan yang direpresentasikan dalam *embedding*.

1.2.4 IndoBERTweet

IndoBERTweet merupakan model *language representation* berbasis *Transformer* yang dikembangkan khusus untuk memahami teks berbahasa Indonesia yang bersifat informal dari media sosial, terutama Twitter. Tujuan utama IndoBERTweet adalah untuk menangkap karakteristik unik bahasa Indonesia di media sosial, seperti penggunaan singkatan, slang, emotikon, hingga struktur kalimat informal yang tidak ditemukan pada teks formal. Secara arsitektural, IndoBERTweet mengadopsi struktur *Transformer encoder* seperti pada BERT, yang terdiri atas lapisan *self-attention* untuk memahami konteks kata secara bidireksional. Dengan pendekatan ini, model dapat merepresentasikan makna suatu kata berdasarkan konteks di sekitarnya, bukan hanya secara urutan kiri-kanan seperti pada model konvensional. Berdasarkan penelitian yang dilakukan Indriani et al. (2023), IndoBERTweet lebih unggul dibanding model bahasa Indonesia seperti IndoBERT dalam memahami gaya bahasa alami media sosial, yang sering kali bersifat informal.

1.2.5 Stratified K-Fold Cross Validation

Stratified K-Fold Cross Validation merupakan salah satu teknik evaluasi model yang digunakan untuk mengukur kinerja algoritma *machine learning*, khususnya pada permasalahan klasifikasi (GeeksforGeeks, 2025). Teknik ini dikembangkan sebagai penyempurnaan dari metode *K-Fold Cross Validation* konvensional dengan menambahkan mekanisme *stratification* (Widodo et al., 2022). Pada metode ini, dataset dibagi ke dalam K lipatan (*fold*) sedemikian rupa sehingga proporsi kelas pada setiap fold merepresentasikan distribusi kelas pada dataset asli. Dengan demikian, setiap fold memiliki komposisi kelas yang relatif seimbang dan tidak menyimpang dari distribusi data sebenarnya (Allen et al., 2021).

Menurut Szeghalmy & Fazekas (2023), keunggulan utama *Stratified K-Fold Cross Validation* terletak pada kemampuannya dalam menangani permasalahan data tidak seimbang (*imbalanced data*), yang umum ditemukan pada tugas klasifikasi dunia nyata. Tanpa stratifikasi, pembagian data secara acak berpotensi menghasilkan fold yang didominasi oleh satu kelas tertentu, sehingga evaluasi model menjadi bias dan kurang representatif. Dengan mempertahankan distribusi kelas pada setiap *fold*, *Stratified K-Fold Cross Validation* mampu menghasilkan estimasi performa model yang lebih stabil, adil, dan reliabel, terutama dalam mengukur metrik evaluasi pada masing-masing kelas.

Gambar 2. Ilustrasi Mekanisme *Stratified K-Fold Cross Validation* (Duan, 2023)

Sebagai ilustrasi stratified cross validation pada Gambar 2, misalkan sebuah dataset klasifikasi terdiri dari dua kelas dengan distribusi tidak seimbang, yaitu 80% data kelas mayoritas dan 20% data kelas minoritas. Pada Stratified K-Fold Cross Validation dengan nilai *K = 5*, dataset akan dibagi menjadi lima *fold* dengan tetap mempertahankan rasio kelas 80:20 pada setiap *fold*. Selanjutnya, pada setiap iterasi, satu *fold* digunakan sebagai data pengujian, sementara empat *fold* lainnya digunakan sebagai data pelatihan. Proses ini diulang hingga seluruh *fold* pernah berperan sebagai data uji, dan hasil evaluasi dari setiap iterasi kemudian dirata-ratakan untuk memperoleh estimasi kinerja model yang lebih stabil dan representatif terhadap distribusi data sebenarnya, serta dihitung nilai standar deviasi untuk mengukur tingkat variasi performa model antar fold.

1.2.6 Blackbox Testing

Blackbox testing merupakan metode pengujian perangkat lunak yang berfokus pada pengujian fungsionalitas sistem tanpa memperhatikan struktur internal, logika program, maupun implementasi kode sumber (Agrawal, 2025). Pada metode ini, penguji hanya memperhatikan kesesuaian antara *input* yang diberikan dan *output* yang dihasilkan berdasarkan spesifikasi kebutuhan sistem. Teknik ini umum digunakan untuk menguji aspek seperti validasi data, alur proses, antarmuka pengguna, serta respons sistem terhadap kesalahan. Karena tidak memerlukan pengetahuan teknis mengenai kode program, *blackbox testing* sangat efektif digunakan untuk mengevaluasi sistem dari sudut pandang pengguna dan memastikan bahwa sistem memenuhi kebutuhan fungsional yang telah ditetapkan.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka peneliti merumuskan masalah sebagai berikut.

1. Bagaimana mengembangkan dan mengukur performa sistem deteksi komentar promosi judi *online* di platform youtube menggunakan arsitektur kombinasi

IndoBERTweet dan GraphSAGE dengan mempertimbangkan hubungan struktural fitur semantik dan *behavioral*?
2. Bagaimana mengimplementasikan model yang dikembangkan ke *web application* untuk mengurangi penyebaran promosi judi *online* di platform youtube?

1.4 Tujuan dan Manfaat

1.4.1 Tujuan

Berdasarkan rumusan masalah, tujuan dari penelitian ini adalah sebagai berikut.

1. Mengembangkan dan mengukur performa sistem deteksi komentar promosi judi *online* di platform youtube menggunakan arsitektur kombinasi IndoBERTweet dan GraphSAGE dengan mempertimbangkan fitur semantik dan *behavioral*.
2. Mengimplementasikan model yang dikembangkan ke *web application* untuk mengurangi penyebaran promosi judi *online* di platform youtube.

1.4.2 Manfaat

Adapun manfaat yang dapat diperoleh dari penelitian yang dilakukan adalah sebagai berikut:

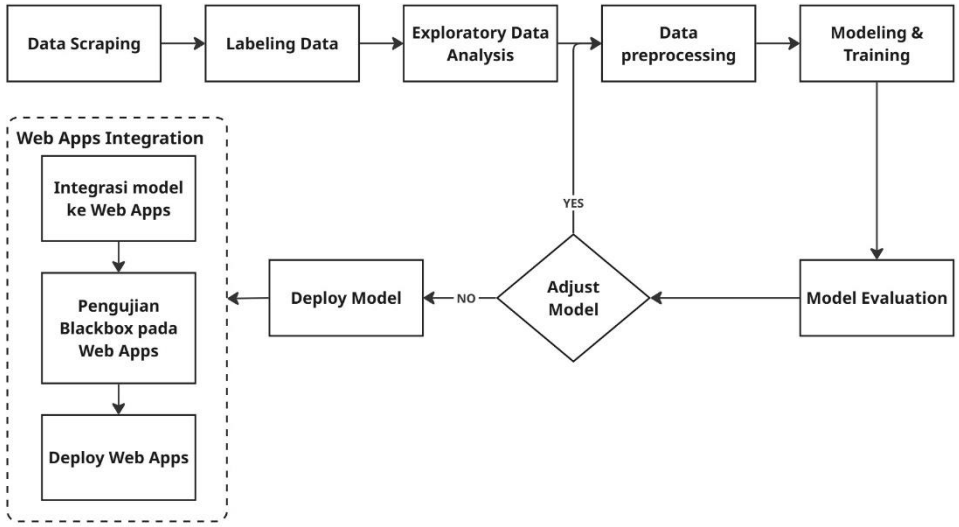
1. Memberikan kontribusi ilmiah baru dalam menggabungkan fitur semantik dan perilaku pengguna melalui pendekatan *hybrid* untuk mendeteksi spam di media sosial.
2. Memberikan solusi nyata dalam membantu platform YouTube, peneliti, dan pemangku kebijakan untuk mendeteksi dan menekan penyebaran komentar yang mempromosikan judi *online*.

BAB II

METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini akan dilakukan secara bertahap mulai dari studi literatur hingga pada tahap integrasi model pada aplikasi web. Tahapan penelitian tersebut ditunjukkan pada Gambar 3.



Gambar 3. Tahapan penelitian yang dilakukan.

2.1.1 Data Scraping

Dataset yang digunakan dalam penelitian ini diperoleh melalui proses *scraping* komentar dari sejumlah video di platform YouTube. Pengumpulan data dilakukan dengan memanfaatkan YouTube API v3, yang memungkinkan pengambilan informasi komentar serta metadata terkait aktivitas pengguna. Data hasil scraping selanjutnya akan digunakan sebagai dasar dalam pembentukan *graph* dan pelatihan model deteksi komentar promosi judi online.

2.1.2 Labeling Data

Dataset yang telah dikumpulkan akan dilabel untuk menandai komentar youtube yang mengandung promosi judi *online*. Komentar yang mengandung promosi judi *online* akan dilabel dengan angka 1 dan komentar biasa akan dilabel menggunakan angka 0. Proses pelabelan akan dilakukan secara manual oleh penulis dengan meninjau setiap komentar secara kontekstual guna mengidentifikasi kecenderungannya melakukan promosi judi *online* atau hanya komentar biasa.

2.1.3 *Exploratory Data Analysis* (EDA)

Tahap ini dilakukan untuk memperoleh pemahaman awal terhadap karakteristik dan pola yang terdapat pada data komentar YouTube. Analisis ini bertujuan untuk mengidentifikasi pola semantik dari komentar dan kecenderungan perilaku pengguna yang berkomentar. Hasil analisis tersebut akan digunakan sebagai dasar dalam proses pembersihan data dan penentuan fitur yang akan digunakan untuk membangun model.

2.1.4 Data preprocessing

Data yang telah dilabel akan diproses dalam beberapa langkah sebelum digunakan pada proses membuat model. Tahapan tersebut terdiri dari *username removal*, menormalkan unicode, penghapusan *timestamp* menit, *punctuation removal*, *unicode normalization*, *emoji normalization*, dan *lowercasing*.

Username Removal. Pada komentar YouTube, tanda @ digunakan untuk menyebut atau menandai pengguna lain dalam komentar (reply). Bagian *username* tersebut akan dihapus karena tidak mengandung informasi semantik yang relevan dengan konten komentar, serta berpotensi mengganggu proses analisis teks.

Timestamp Removal. Banyak pengguna YouTube menuliskan *timestamp* berupa penanda waktu dalam format menit-detik (misalnya *"1:23"*, *"05:10"*, atau *"12.45"*) untuk merujuk pada momen tertentu di dalam video. Namun, pada konteks komentar promosi judi online, pola penulisan *timestamp* sering digunakan secara manipulatif tanpa merujuk pada isi video, melainkan sebagai strategi penyamaran agar terlihat seperti komentar organik dari penonton. Contoh komentar judi yang menggunakan timestamp pada komentarnya dapat dilihat pada berikut.

08:17 Bener-bener niat bikin kontennya **BERKAH99**
19:27 Transisinya smooth! **BERKAH99**
20:20 Pantas trending! **BERKAH99**
19:45 Kualitas premium! **BERKAH99**
19:28 Nyaman banget ditonton! **BERKAH99**
15:23 Konsistennya bikin salut! **BERKAH99**
17:33 Udah keren, gak perlu banyak efek! **BERKAH99**
19:33 Selalu menarik! **BERKAH99**
06:27 Suara dan visualnya sinkron banget! **BERKAH99**
05:22 Selalu dinanti nih kontennya! **BERKAH99**
05:59 Sumpah keren parah! **BERKAH99**
14:15 Wajib viral ini!**BERKAH99**
19:23 Kualitas premium! **BERKAH99**

Gambar 4. Contoh komentar promosi judi yang menggunakan timestamp

Punctuation Removal. Seluruh tanda baca pada teks komentar akan dihapus pada tahap *preprocessing*. Tanda baca tidak membawa informasi semantik yang signifikan untuk proses klasifikasi, dan keberadaannya berpotensi menimbulkan *noise* pada proses pembentukan representasi teks. Dengan menghilangkan tanda baca, struktur teks menjadi lebih bersih dan konsisten, sehingga dapat membantu model dalam memfokuskan pembelajaran pada konten kata yang bermakna.

Unicode Normalization. Komentar YouTube sering mengandung karakter *Unicode non-standar*, seperti huruf beraksen, simbol dekoratif, maupun karakter dari

berbagai sistem tulisan, yang berpotensi mengganggu proses tokenisasi dan ekstraksi fitur. Unicode non-standar dalam penelitian ini didefinisikan sebagai karakter yang berada di luar rentang *Basic Latin* (U+0000–U+007F). Rentang *Basic Latin* mencakup huruf alfabet (A–Z), angka (0–9), tanda baca, serta karakter kontrol seperti DELETE (Unicode Consortium, 2025). Karakter di luar rentang tersebut kerap dimanfaatkan oleh pelaku promosi judi *online* untuk menghindari sistem deteksi otomatis, karena tokenisasi akan memperlakukan karakter yang secara visual serupa tetapi berbeda representasi *unicode* sebagai token yang berbeda (misalnya huruf *a* dan *á*). Oleh karena itu, proses normalisasi *unicode* diterapkan untuk menyamakan representasi karakter sehingga konsistensi token dan kualitas fitur teks dapat terjaga. Normalisasi dilakukan menggunakan pustaka AnyASCII untuk mengubah seluruh karakter *unicode* ke dalam format ASCII yang setara. Misalnya huruf beraksen (á, é, ü) akan diubah menjadi format ASCII (a, e, u). Normalisasi ini membantu menyederhanakan representasi teks, mengurangi variasi karakter yang tidak relevan, dan meningkatkan konsistensi data sehingga model dapat melakukan pembelajaran secara lebih efektif.

Lowercasing. Seluruh teks pada komentar akan diubah menjadi huruf kecil (*lowercase*) untuk memastikan konsistensi penulisan kata. Perbedaan kapitalisasi tidak memiliki makna semantik yang berbeda dalam konteks klasifikasi komentar, tetapi dapat dianggap sebagai token berbeda oleh model. Proses *lowercasing* membantu mengurangi jumlah variasi kata yang tidak perlu, sehingga dapat meningkatkan kualitas representasi teks dan efisiensi proses pembelajaran.

Emoji Normalization. Proses normalisasi emoji dilakukan dengan memanfaatkan library Emoji untuk mengonversi setiap emoji ke dalam representasi tekstual. Pendekatan ini merujuk pada pedoman yang digunakan dalam penelitian IndoBERTweet oleh Koto et al., di mana IndoBERTweet telah dilatih untuk memahami makna emoji dalam konteks semantik. Dengan demikian, normalisasi emoji diterapkan untuk mengevaluasi sejauh mana informasi atau pola yang terkandung dalam emoji dapat memberikan kontribusi terhadap kemampuan model dalam mempelajari dan mengenali karakteristik komentar.

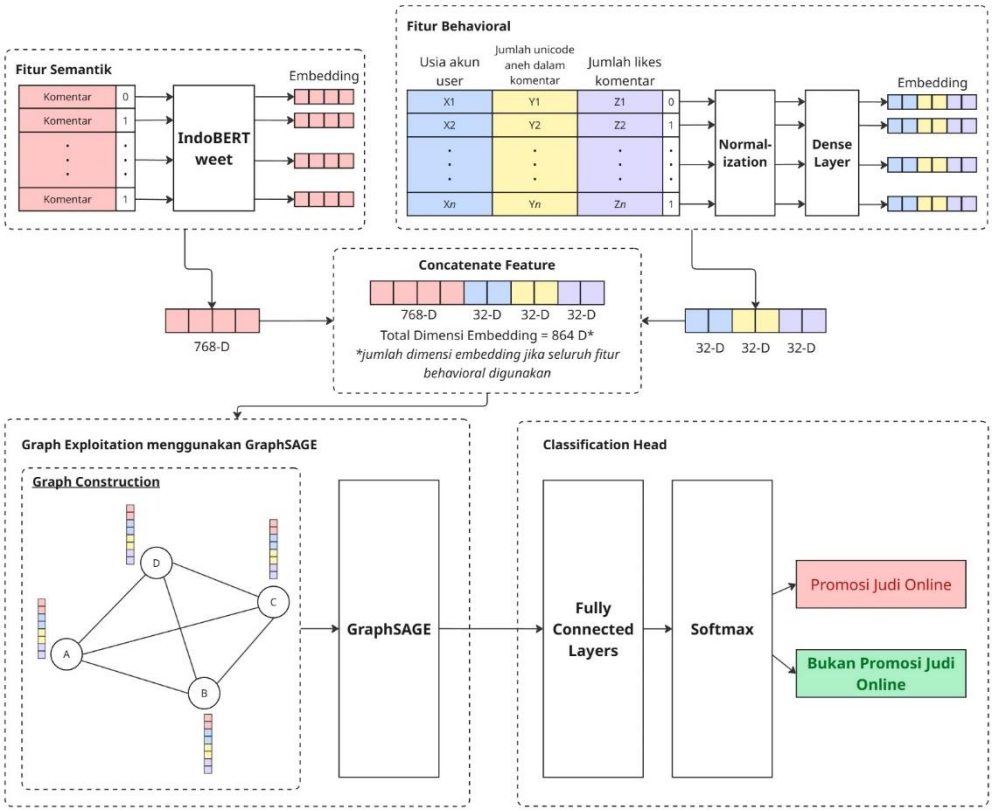
Pada tahap akhir *preprocessing*, dataset akan dibagi menjadi tiga bagian dengan rasio 70:15:15, yaitu 70% sebagai data latih (*training set*), 15% sebagai data validasi (*validation set*), dan 15% sebagai data uji (*testing set*).

2.1.5 Modeling & Training

Penelitian ini menggunakan arsitektur kombinasi IndoBERTweet-GraphSAGE untuk menganalisis pola semantik dan perilaku pengguna berdasarkan *metadata* pada video YouTube. Menurut Zouak et al. (2025), kombinasi arsitektur BERT dan GraphSAGE efektif dalam menangkap konten semantik maupun hubungan struktural antar konten. IndoBERTweet unggul dalam memahami makna teks yang informal secara kontekstual, sementara GraphSAGE, dengan kemampuan *inductive learning*-nya, mampu menggeneralisasi representasi *node* baru tanpa perlu mempelajari graph baru secara keseluruhan. Selain itu, penelitian ini juga mengeksplorasi integrasi fitur *behavioral* pada representasi *node*, sejalan dengan temuan Alhassun & Rassam (2022) yang

menunjukkan bahwa fitur perilaku pengguna memberikan kontribusi dalam meningkatkan akurasi deteksi spam pada komentar berbasis media sosial.

Kombinasi ini memungkinkan deteksi spam yang lebih akurat efisien dalam lingkungan media sosial yang dinamis seperti YouTube, di mana konten dan interaksi pengguna terus berkembang. Model ini juga memiliki keunggulan dalam hal skalabilitas dan adaptabilitas guna meminimalkan beban pelatihan ulang secara menyeluruh pada setiap perubahan data.



Gambar 5. Kombinasi arsitektur IndoBERTweet-GraphSage dalam sistem deteksi

Rancangan arsitektur yang ditunjukkan pada Gambar 5 merupakan representasi dari kombinasi arsitektur IndoBERTweet-GraphSAGE yang digunakan dalam pengembangan sistem deteksi komentar promosi judi *online*. Arsitektur ini dirancang untuk mengintegrasikan analisis semantik berbasis teks dan perilaku pengguna dengan pemodelan relasi struktural. Secara keseluruhan, kombinasi kedua komponen tersebut menjadi dasar utama dalam membangun model deteksi yang lebih komprehensif, sebagaimana akan dijelaskan secara lebih rinci pada penjelasan berikut.

Fitur Semantik. Komentar dari video YouTube akan diolah menggunakan IndoBERTweet untuk mengekstraksi representasi semantik dari setiap komentar sebagaimana yang dilakukan oleh Zouak et al. dalam penelitian deteksi spam-nya. Komentar akan melakui proses tokenisasi terlebih dahulu. Setiap token pada komentar akan dipetakan dalam *embedding* berdimensi 768. Kemudian, *embedding* dari seluruh

token dalam komentar akan di rata-ratakan dengan mekanisme *mean pooling* untuk merangkum semua token *embedding* menjadi satu vektor tetap per komentar. Representasi ini dilakukan agar dapat mencerminkan makna kontekstual dan pola bahasa alami dalam teks komentar.

Fitur *Behavioral*. Pola perilaku yang diperoleh dari metadata pengguna akan melalui proses normalisasi terlebih dahulu menggunakan MinMaxScaler pada library python scikit-learn. Normalisasi MinMaxScaler diperlukan karena fitur *behavioral* memiliki rentang nilai yang sangat bervariasi, misalnya jumlah *likes* yang dapat mencapai ribuan dibandingkan dengan jumlah karakter *unicode* yang umumnya hanya berkisar puluhan. Perbedaan skala ini dapat menyebabkan fitur dengan nilai besar mendominasi proses pembelajaran, sementara fitur lain yang juga relevan justru terabaikan (Alhassun & Rassam, 2022). Dengan MinMaxScaler, seluruh fitur dinyatakan dalam rentang 0–1 sehingga model dapat memprosesnya secara seimbang, lebih cepat, dan menghasilkan akurasi yang lebih baik (Montminy et al., 2013). Setelah dinormalisasi, fitur *behavioral* akan diproses pada *dense layer* sebelum digabung dengan *embedding* semantik. *Dense layer* berfungsi sebagai mekanisme pemetaan agar fitur numerik dapat diubah menjadi representasi vektor yang selaras dengan ruang *embedding* semantik (Alom et al., 2020). Proses pembentukan embedding pada dense layer dilakukan menggunakan library PyTorch melalui fungsi torch.nn.Linear, yang merepresentasikan transformasi linear terhadap fitur masukan. Secara umum, operasi yang dilakukan oleh lapisan ini dapat dinyatakan pada persamaan (3)

$$y = Wx + \mathrm{b} \tag{3}$$

dengan x merupakan vektor *input features* $(in\_feature)$, W adalah bobot, dan b adalah bias. Pada tahap inisialisasi awal sebelum proses pelatihan, parameter bobot dan bias tidak ditentukan secara deterministik, melainkan diinisialisasi secara acak mengikuti distribusi uniform. Skema inisialisasi yang digunakan oleh torch.nn.Linear dapat dirumuskan sebagaimana rumus

$$W_{ij} \sim U(-\sqrt{k}, \sqrt{k}) \tag{4}$$

$$b_i \sim U(-\sqrt{k}, \sqrt{k}) \tag{5}$$

dengan:

$$k = \frac{1}{in\_feature} \tag{6}$$

Pada persamaan tersebut, W_{ij} menyatakan elemen bobot pada baris ke-i dan kolom ke-j dari matriks bobot W, sedangkan $b_i$merepresentasikan elemen bias ke-i. Simbol $U\ (-\sqrt{k}, \sqrt{k})$ menunjukkan distribusi uniform dengan rentang nilai antara $-\sqrt{k}$ dan $\sqrt{k}$. Parameter in_features menunjukkan jumlah fitur masukan pada lapisan linear.

Skema inisialisasi menggunakan distribusi uniform dengan membatasi nilai awal bobot dan bias dalam rentang nilai bertujuan untuk menjaga kestabilan varians aktivasi pada lapisan awal jaringan. Dengan demikian, risiko terjadinya *vanishing* maupun *exploding activation* dapat diminimalkan pada tahap awal pelatihan (Glorot & Bengio, 2010). Meskipun nilai parameter bersifat acak pada awal proses, bobot dan bias akan diperbarui secara iteratif melalui mekanisme *backpropagation*, sehingga embedding

yang dihasilkan secara bertahap menjadi semakin representatif terhadap pola data dan label yang dipelajari oleh model.

Menurut Wu et al. (2024), fitur numerik yang secara alami berbentuk nilai skalar memiliki keterbatasan apabila digunakan secara langsung atau hanya melalui penskalaan linear sederhana. Peneliti tersebut menjelaskan bahwa *pendekatan linearly-scaled embedding*, yaitu mengalikan satu nilai skalar dengan bobot *embedding* secara linear, sering kali tidak mampu merepresentasikan informasi secara kaya. Hal ini disebabkan karena seluruh karakteristik fitur numerik dipadatkan hanya ke dalam satu nilai tunggal, sehingga informasi yang dapat dipelajari oleh jaringan saraf menjadi terbatas. Kondisi ini disebut sebagai *information bottleneck*, yaitu situasi ketika representasi input tidak cukup ekspresif untuk dimanfaatkan secara optimal oleh model jaringan saraf yang kompleks, sehingga model kesulitan menangkap pola laten maupun hubungan non-linear yang penting. Oleh karena itu, akan dilakukan pembentukan *feature embedding* dari nilai skalar pada fitur *behavioral*.

Peneliti Liu et al. (2020) dan Kang et al. (2021) juga menyoroti bahwa penggunaan *embedding* dengan ukuran yang terlalu besar dapat menyebabkan dominasi representasi tertentu dalam model, terutama pada fitur dengan kardinalitas tinggi. Kondisi ini berpotensi membuat fitur lain menjadi kurang berkontribusi dalam proses pembelajaran. Prinsip serupa juga relevan dalam integrasi *embedding* teks dan fitur perilaku, di mana *embedding* teks yang berdimensi tinggi berisiko mendominasi informasi dari fitur numerik. Oleh karena itu, dalam penelitian ini, fitur perilaku direpresentasikan dalam bentuk *embedding* berdimensi tetap sebanyak 32 dimensi per fitur. Pendekatan ini bertujuan menjaga keseimbangan representasi sehingga informasi perilaku tidak kalah oleh dominasi *embedding* teks, namun tetap cukup ekspresif untuk menangkap karakteristik perilaku pengguna.

Dalam penelitian ini, fitur perilaku (*behavioral*) yang digunakan meliputi usia akun pengguna, jumlah likes pada komentar, dan jumlah karakter tidak lazim dalam komentar. Setiap fitur perilaku direpresentasikan dalam bentuk *embedding* berdimensi 32 untuk menjaga keseimbangan kontribusi antara fitur perilaku dan tidak mendominasi *embedding* teks yang menjadi fitur utama. Dengan demikian, apabila ketiga fitur perilaku tersebut digunakan secara bersamaan, maka total dimensi *embedding* perilaku adalah 96 dimensi (3 × 32).

***Concatenate Feature*.** Hasil representasi dari fitur semantik dan fitur *behavioral* akan digabungkan (*concatenate*) menjadi satu vektor fitur gabungan. Vektor ini merepresentasikan baik sisi konten (semantik) maupun sisi perilaku (*behavioral*) dari komentar. Saat fitur semantik dari IndoBERTweet (768 dimensi) digabungkan dengan fitur perilaku pengguna, jumlah dimensi *embedding* akan bertambah sesuai jumlah fitur perilaku yang digunakan. Misalnya, jika 3 jenis fitur perilaku digunakan yang masing-masing berjumlah 32 dimensi, maka *embedding* akhir akan memiliki 768 + (32 × 3) = 864 dimensi.

Graph Construction. Pada tahap ini, struktur graf dibangun dari data sebelum digunakan dalam proses pembelajaran menggunakan GraphSAGE. Struktur graf terdiri dari *node* dan *edge*. Pada eksperimen awal, konstruksi graf tidak lagi hanya mengandalkan relasi kesamaan semantik, melainkan dirancang berdasarkan integrasi fitur semantik dan perilaku pengguna yang diperoleh dari hasil eksplorasi data. Setiap

node merepresentasikan satu komentar dengan fitur awal berupa kombinasi lengkap yang mencakup *embedding* komentar hasil pemrosesan IndoBERTweet, usia akun pengguna, jumlah karakter tidak lazim, serta jumlah likes pada komentar. *Edge* antar *node* dibentuk menggunakan kombinasi seluruh pola relasi yang teridentifikasi dari hasil *exploratory data analysis*, mencakup kesamaan semantik, kedekatan temporal, serta kesamaan pola aktivitas pengguna. Pendekatan ini digunakan sebagai konfigurasi dasar untuk mengevaluasi kemampuan model dalam memanfaatkan relasi struktural dan perilaku pengguna secara komprehensif pada tugas deteksi komentar promosi judi online.

Graph Exploitation. *Graph exploitation* merupakan tahap pemanfaatan struktur graf untuk mengungkap pola relasi antar *node* atau entitas yang telah dibentuk pada tahap konstruksi graf. Pada tahap ini, model GraphSAGE memanfaatkan informasi dari *node* tetangga melalui proses agregasi untuk menghasilkan representasi yang lebih informatif bagi setiap *node*. Dengan demikian, model tidak hanya belajar dari fitur individual setiap komentar, tetapi juga dari hubungan kontekstual yang terbentuk di dalam graf, seperti kesamaan konten atau keterkaitan perilaku pengguna. Hasil dari tahap ini diharapkan mampu menangkap pola relasional yang relevan dalam mendeteksi komentar promosi judi *online* secara lebih akurat.

Classification Head. Hasil representasi *node* dari GraphSAGE akan diteruskan ke *fully connected layers* untuk menghasilkan skor (logit) terhadap setiap kelas yang ditargetkan. Setelah itu, skor tersebut akan melalui fungsi *softmax* untuk mengubahnya menjadi probabilitas klasifikasi akhir.

Proses *training* awal dilakukan dengan fitur, *edge*s, dan parameter *default* yang ditentukan berdasarkan hasil EDA dan hasil studi literatur pada penelitian Zouak et al. (2025), Alhassun & Rassam (2022), dan Hamilton et al. (2017).

Tabel 1. *Hyperparameter default* yang digunakan untuk pelatihan awal model

No.	Eksperimen	Nilai
1	Batch Size	128
2	Epoch	50
3	Aggregator	Mean
4	Learning rate	0.001

Tabel 2. Fitur dan *Edge*s *default* yang digunakan untuk pelatihan awal model

Urutan	Eksperimen	Status
1	Preprocessing Penggunaan Emoji	Digunakan
2	Preprocessing Hapus Menit	Digunakan
3	Fitur Usia Akun	Digunakan
4	Fitur Jumlah Like	Digunakan
5	Fitur Jumlah Karakter Tidak Lazim	Digunakan
6	*Edge*s Semua Komentar User di Video yang Sama	Digunakan
7	*Edge*s Komentar yang Usernya berkomentar > 1 kali	Digunakan

Urutan	Eksperimen	Status
8	*Edge*s Komentar User yang Waktu Uploadnya Berdekatan dalam Rentang Waktu <= 5 menit	Digunakan
9	*Edge*s Komentar dengan cosine similarity >= 0.7	Digunakan
10	*Edge*s Komentar yang Jumlah Karakter Tidak Lazim > 5	Digunakan

2.1.6 *Model Evaluation*

Model yang telah dilatih selanjutnya dievaluasi untuk menilai kemampuannya dalam mengklasifikasikan komentar dengan tepat. Evaluasi ini dilakukan guna mengukur sejauh mana model mampu mengenali pola dari data secara akurat. Kinerja model kemudian dinilai menggunakan metrik umum pada tugas klasifikasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Dalam penelitian ini, metriks umum tersebut diinterpretasikan dengan makna seperti berikut.

Accuracy. Metriks ini merupakan persentase prediksi yang benar dari semua komentar yang diuji. Metriks ini dinyatakan dalam persamaan (7)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Dimana:

- *True Positive* (TP): jumlah prediksi benar komentar promosi judi *online* (kelas positif)
- *True Negative* (TN): jumlah prediksi benar komentar biasa (kelas negatif)
- *False Positive* (FP): jumlah prediksi salah komentar promosi judi *online* (kelas positif)
- *False Negative* (FN): jumlah prediksi salah komentar biasa (kelas negatif)

Precision. Metrik ini digunakan untuk mengukur ketepatan model dalam memprediksi komentar promosi judi. Interpretasi dalam penelitian ini berarti semakin tinggi *precision*, semakin jarang model salah menandai komentar biasa sebagai promosi judi. Metriks ini dinyatakan dalam persamaan (8)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

Recall. Metrik ini digunakan untuk mengukur proporsi komentar promosi judi yang berhasil terdeteksi oleh model. Interpretasi dalam penelitian ini berarti semakin tinggi *recall*, semakin sedikit komentar promosi judi yang dilewatkan atau tidak terdeteksi. Metriks ini dinyatakan dalam persamaan (9)

$$\text{Precision} = \frac{TP}{TP + FN} \tag{9}$$

F1-score. Metrik ini digunakan untuk mengukur keseimbangan antara *precision* dan *recall*, sehingga menunjukkan seberapa baik model mampu mendeteksi komentar promosi judi tanpa terlalu banyak melakukan kesalahan positif maupun negatif.

Interpretasi dalam penelitian ini, semakin tinggi *f1-score* berarti model baik dalam mendeteksi komentar promosi judi tanpa terlalu banyak salah baik kesalahan positif maupun negatif. Metriks ini dinyatakan dalam persamaan (10)

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

2.1.7 Adjust Model

Pada tahap ini, performa model akan dioptimalkan berdasarkan hasil *testing* dan *evaluation*. Model akan melalui proses *reverse ablation study* untuk melihat kontribusi dari masing-masing fitur dan *edge* yang digunakan pada model. Metode ini dimulai dengan cara mengaktifkan seluruh fitur dan seluruh tipe *edge* pada model, kemudian menonaktifkan satu komponen pada satu waktu untuk mengamati dampaknya terhadap performa model.

Jika menonaktifkan suatu fitur atau *edge* menghasilkan penurunan performa, maka komponen tersebut dianggap memiliki kontribusi positif terhadap kemampuan model dalam melakukan klasifikasi. Sebaliknya, jika menonaktifkan suatu fitur atau *edge* justru menghasilkan peningkatan performa, maka komponen tersebut dianggap sebagai *noise* yang mengganggu proses pembelajaran model, sehingga dapat dihapus dari konfigurasi pelatihan. Metode *reverse ablation study* umum digunakan dalam penelitian *Graph Neural Network* (GNN) sebagaimana yang dilakukan Luo et al. (2025) karena kinerja model sangat bergantung pada informasi struktural, khususnya hubungan antar *node* dan *edge* di dalam *graph* (Yuan et al., 2023). Dengan demikian, penilaian kontribusi tiap komponen tidak dapat dilakukan tanpa mempertimbangkan struktur *graph* secara utuh.

Dalam penelitian ini, model akan di-*adjust* menggunakan beberapa skenario eksperimen, yaitu berdasarkan teknik *preprocessing*, fitur *behavioral* pengguna, dan pemilihan *edge*s pada *graph construction*. Pendekatan ini dilakukan untuk memperoleh konfigurasi model yang paling optimal dan memiliki generalisasi yang baik dalam deteksi promosi judi online. Skenario tersebut dijelaskan lebih rinci seperti berikut.

Skenario Teknik *Preprocessing*. Pada skenario ini akan dilakukan uji coba terhadap pengaruh emoji *removal* dan penghapusan timestamp menit terhadap akurasi pada model. Langkah ini bertujuan untuk mengetahui sejauh mana keberadaan emoji dalam teks komentar memengaruhi hasil klasifikasi. Dalam konteks promosi judi *online*, emoji kerap dimanfaatkan untuk menarik perhatian atau menghindari deteksi otomatis oleh sistem moderasi. Selain itu, untuk timestamp *removal* akan divalidasi apakah benar menit tersebut hanya *noise* atau ternyata memiliki pola yang bisa dipelajari.

Skenario Fitur *Behavioral*. Pada skenario ini dilakukan uji coba pemilihan fitur *behavioral* untuk menganalisis pengaruh karakteristik perilaku pengguna terhadap performa model deteksi komentar promosi judi *online*. Fitur *behavioral* yang dimaksud mencerminkan pola aktivitas pengguna, seperti usia akun, jumlah *likes* pada komentar, serta jumlah karakter tidak lazim pada komentar. Melalui pengujian ini, dilakukan evaluasi terhadap kombinasi fitur mana yang paling berkontribusi dalam membedakan komentar promosi dengan komentar normal.

Skenario Pemilihan *Edge*s pada *Graph* Construction. Pada skenario ini dilakukan uji coba pemilihan jenis *edges* pada tahap *graph construction* untuk mengetahui pengaruh masing-masing relasi terhadap kinerja model GraphSAGE dalam mendeteksi komentar promosi judi *online*. Setiap jenis *edge* (Tabel 2) merepresentasikan bentuk hubungan yang berbeda antar komentar, seperti hubungan berdasarkan pengguna (*user-based edges*), kedekatan waktu publikasi komentar (*temporal edges*), serta kesamaan konten secara semantik (*semantic similarity edges*). Melalui pengujian ini, dilakukan analisis terhadap kombinasi *edges* yang paling efektif dalam memperkuat representasi struktural graf.

Hasil terbaik dari skenario *adjust* model selanjutnya dianalisis secara lebih mendalam menggunakan dua pendekatan evaluasi. Pendekatan pertama dilakukan dengan menguji model terbaik pada *unseen* data yang diperoleh melalui proses scraping setelah data pelatihan. Analisis pertama dilakukan guna mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

Pendekatan kedua bertujuan untuk menganalisis pengaruh ketidakseimbangan data pelatihan (*imbalanced data*) terhadap kinerja model. Untuk keperluan ini, digunakan metode *stratified 5-fold cross validation*, yang memastikan proporsi kelas pada setiap fold tetap merepresentasikan distribusi data asli. Metode *stratified 5-fold cross validation* secara umum digunakan tidak hanya sebagai teknik validasi model, tetapi juga sebagai alat analisis stabilitas dan ketahanan model terhadap variasi pembagian data, khususnya pada kondisi distribusi kelas yang tidak seimbang sebagaimana yang dilakukan ada penelitian (Alonso & Escot, 2025).

Pemilihan 5-fold pada stratified cross validation dalam penelitian ini didasarkan pada pertimbangan keseimbangan antara akurasi estimasi dan efisiensi komputasi. Rodríguez et al. (2010) menyatakan bahwa estimasi performa model dengan jumlah fold 5 atau 10 cenderung menghasilkan bias yang lebih kecil dibandingkan penggunaan jumlah fold yang terlalu sedikit atau terlalu besar. Namun demikian, peningkatan jumlah fold juga berdampak pada meningkatnya biaya komputasi, karena proses pelatihan dan pengujian model harus diulang sebanyak jumlah fold yang digunakan. Wong (2015) menegaskan bahwa ketika nilai k semakin besar, biaya komputasi dari k-fold cross validation akan meningkat secara signifikan. Oleh karena itu, penggunaan 5-fold cross validation dipilih sebagai konfigurasi yang mampu memberikan estimasi performa yang stabil dengan bias yang relatif rendah, sekaligus tetap menjaga efisiensi komputasi, terutama mengingat kompleksitas model dan ukuran data yang digunakan dalam penelitian ini.

2.1.8 Web App Integration

Model yang telah dievaluasi akan di-*deploy* untuk digunakan dalam lingkungan *web application*. *Web application* akan dirancang untuk mendukung proses pencegahan penyebaran promosi judi *online* di platform youtube. Pengguna terbagi atas tiga *roles* secara garis besar, yaitu pemilik video youtube, moderator dari pemilik akun youtube, dan penonton dari video youtube.

2.2 Waktu dan Lokasi Penelitian

Penelitian mulai berlangsung sejak 11 Juli 2025 hingga 29 November 2025. Penelitian ini dilakukan di Lab Artificial Intelligence and Multimedia Processing (AIMP), Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin, Gowa, Sulawesi Selatan.

2.3 Instrumen Penelitian

Adapun beberapa instrumen penelitian yang digunakan dalam penelitian ini, yaitu :

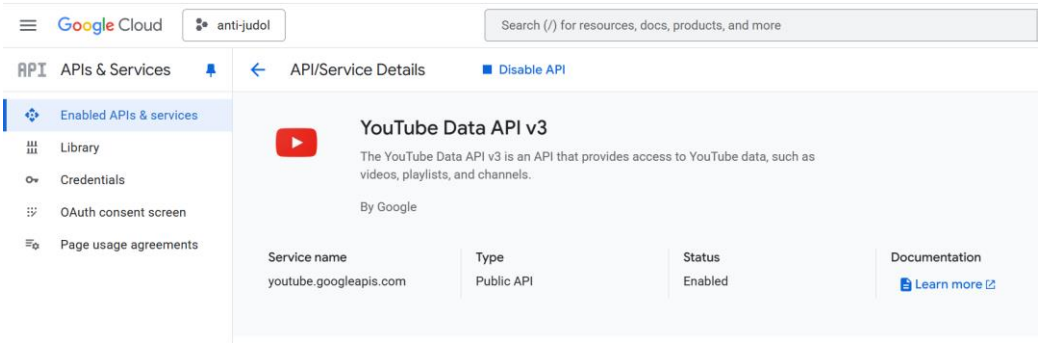
1. Software:
   a. Kaggle Notebook
   b. Visual Studio Code
   c. Postman
2. Bahasa Program:
   a. Python 3.11.13
   b. Javascript
3. Library (AI Model):
   a. Transformers 4.52.4
   b. Torch 2.6.0+cu124
   c. Tqdm 4.67.1
   d. Scikit-learn 1.2.2
   e. Torch-scatter 2.1.2+pt26cu124
   f. Torch-sparse 0.6.18+pt26cu124
   g. Torch-cluster 1.6.3+pt26cu124
   h. Torch-spline-conv 1.2.2+pt26cu124
   i. Torch-geometric 2.6.1
   j. Pandas 2.2.3
   k. Emoji 2.14.1
   l. AnyASCII
4. Library (Web Apps):
   a. FastAPI
   b. Supabase
   c. Next JS
   d. Dotenv
   e. Axios
   f. Tailwind css
   g. googleapiclient
5. Hardware:
   Lenovo Legion 5 CPU Intel Core i7-12700H 20 CPUs 2,7 GHz, GPU Nvidia GeForce RTX 3060 Laptop 6 GB VRAM, 16 GB RAM & SSD 1 TB

2.4 Teknik Pengambilan Data

Dataset yang digunakan dalam penelitian ini diperoleh melalui proses *scraping* komentar dari platform YouTube. Proses pengambilan data dilakukan dengan memanfaatkan YouTube API v3, yang menyediakan akses terhadap berbagai informasi di platform tersebut. Tahapan pengambilan data dilakukan melalui dua langkah utama. Pertama, peneliti memperoleh akses *Application Programming Interface* (API) dengan mendaftarkan *project* pada Google Cloud Console untuk mendapatkan *API key*. Kedua, dilakukan implementasi proses *scraping* dengan memanfaatkan endpoint YouTube API V3 guna mengambil data komentar dan metadata pengguna. Tahap tersebut dijelaskan lebih lanjut pada penjelasan berikut.

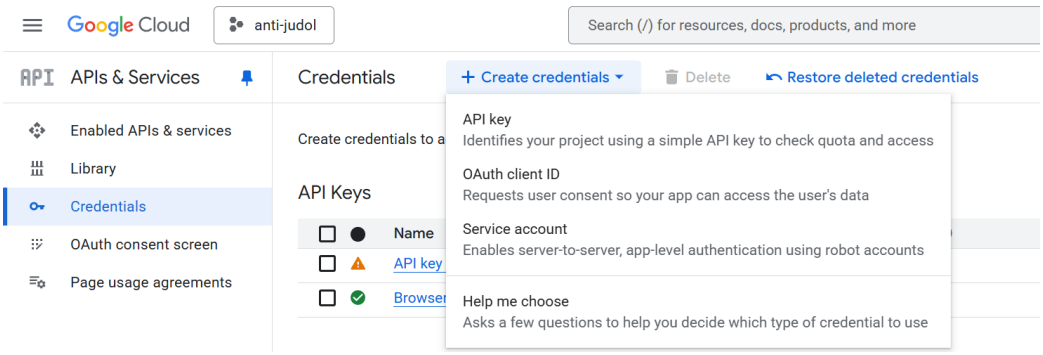
2.4.1 Konfigurasi Akses YouTube API V3

YouTube Data API v3 merupakan layanan yang disediakan oleh Google melalui platform Google Cloud Console. Sebelum layanan ini dapat digunakan, API tersebut harus diaktifkan terlebih dahulu pada konsol Google Cloud. Proses aktivasi dilakukan dengan membuka menu *APIs & Services*, kemudian memilih opsi *Enable APIs & Services*. Selanjutnya, pengguna dapat mencari layanan YouTube Data API v3 dan mengaktifkannya dengan menekan tombol *Enable API* hingga statusnya berubah menjadi *Disable API*, yang menandakan bahwa layanan telah aktif dan siap digunakan.



Gambar 6. Tampilan menu untuk enable layanan Youtube API

Setelah YouTube Data API v3 berhasil diaktifkan, langkah berikutnya adalah memperoleh API key melalui menu *Credentials* pada Google Cloud Console. Proses ini dilakukan dengan menekan tombol *Create Credentials*, kemudian memilih opsi *API Key*. Kunci API yang dihasilkan berfungsi sebagai autentikasi untuk setiap permintaan yang dikirimkan ke layanan YouTube Data API v3.



Gambar 7. Tampilan menu credentials API Key untuk Youtube API

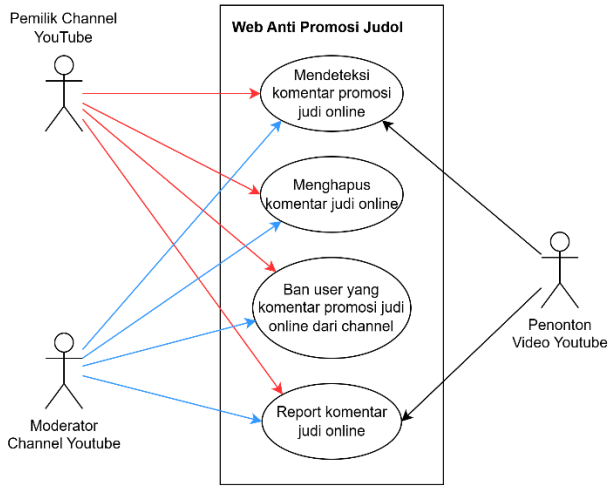
2.4.2 Implementasi Data Scraping di Youtube

Setelah memperoleh *API key*, tahap berikutnya adalah implementasi *data scraping* terhadap data komentar YouTube. Proses ini dilakukan dengan mengirimkan permintaan ke endpoint YouTube Data API v3 menggunakan *API key* yang telah dihasilkan. Melalui API ini, peneliti dapat mengakses data seperti komentar *user*, id komentar, waktu komentar diunggah oleh *user*, *username* yang komentar, id *channel* dari user yang komentar, waktu dibuatnya channel dari *user* yang komentar, dan jumlah *likes* dari komentar.

2.5 Perancangan dan Implementasi Sistem

2.5.1 Use Case Diagram

Use case diagram digunakan untuk menggambarkan hubungan antara aktor dan fungsionalitas utama yang terdapat dalam sistem yang dikembangkan. Pada konteks penelitian ini, *use case diagram* disusun untuk menjelaskan interaksi antara pengguna dan sistem deteksi komentar promosi judi *online.*



Gambar 8. *Use case diagram sistem* deteksi komentar promosi judi *online*

Pemilik *Channel* YouTube. Sebagai pemilik pengguna YouTube, pengguna memiliki kendali penuh terhadap video yang diunggah. Dalam sistem website ini, pemilik pengguna dapat melakukan berbagai tindakan terhadap komentar yang terindikasi mengandung promosi judi *online*, seperti mendeteksi komentar, menghapus komentar tersebut, melakukan *ban* terhadap pengguna yang berkomentar, serta melaporkan komentar terkait kepada pihak YouTube untuk ditindaklanjuti.

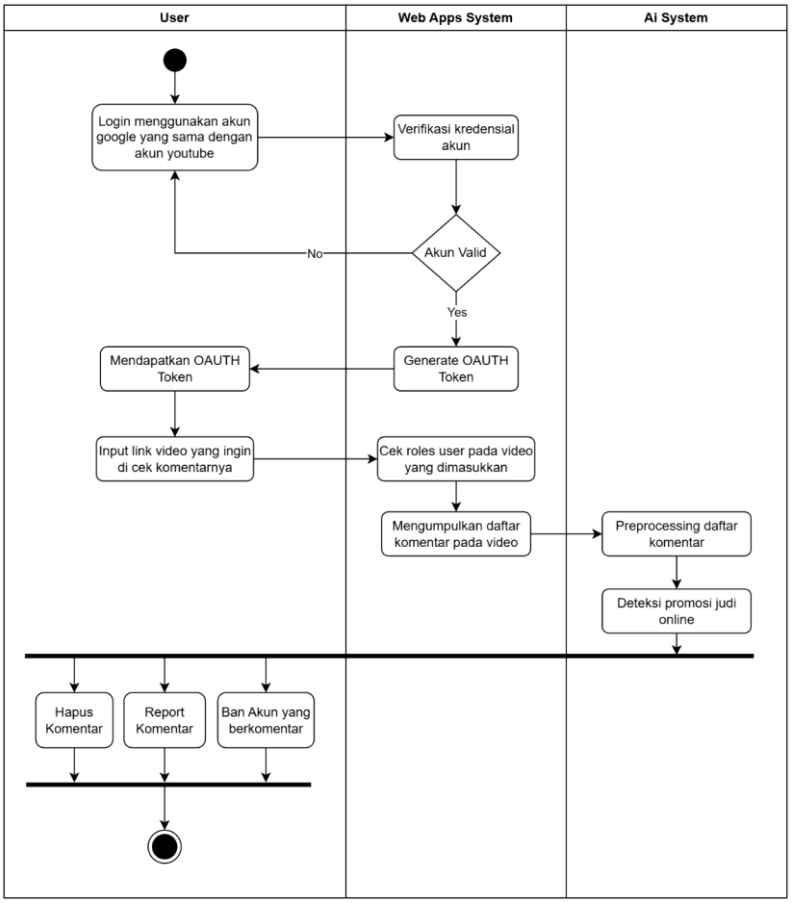
Moderator *Channel* YouTube. Moderator adalah pihak yang dipilih oleh pemilik pengguna akun youtube untuk membantu mengelola pengguna. Moderator dalam konteks ini adalah moderator di platform YouTube yang resmi. Dalam sistem website deteksi komentar promosi judi *online*, peran tersebut dapat dideteksi oleh sistem. Jika akun memiliki akses moderator, maka pengguna dapat melakukan berbagai tindakan terhadap komentar yang terindikasi mengandung promosi judi *online*, seperti mendeteksi komentar promosi, menghapus komentar tersebut, melakukan *ban* terhadap pengguna yang berkomentar, dan melaporkan komentar terkait kepada pihak YouTube untuk ditindaklanjuti.

Penonton Video YouTube. Peran ini memiliki kemampuan untuk melakukan deteksi terhadap komentar yang mengandung unsur promosi judi *online*. Selain itu, pengguna juga dapat melaporkan komentar terkait kepada pihak YouTube untuk ditindaklanjuti lebih lanjut. Partisipasi aktif dari penonton dalam pelaporan komentar bermasalah dapat mempercepat proses penanganan oleh pihak YouTube dan mendukung upaya pencegahan penyebaran konten promosi judi *online*.

2.5.2 Activity Diagram

Activity diagram digunakan untuk menggambarkan alur aktivitas atau proses yang terjadi di dalam sistem secara terperinci. Diagram ini menunjukkan urutan langkah-langkah logis dari setiap proses, mulai dari aktivitas awal hingga tercapainya hasil akhir. Dengan adanya *activity diagram*, aliran kerja sistem dapat dipahami secara lebih jelas,

termasuk bagaimana interaksi antar komponen dan pengambilan keputusan pada setiap tahap proses. Pada penelitian ini, *activity diagram* disusun untuk menjelaskan proses kerja sistem deteksi komentar promosi judi *online*, yang dimulai dari input data komentar, proses klasifikasi oleh model, hingga aksi untuk memoderasi komentar yang telah diklasifikasi. *Activity diagram* sistem deteksi komentar promosi judi online ditunjukkan pada Gambar 9.



Gambar 9. Activity diagram sistem deteksi komentar promosi judi *online*