BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Besarnya skala platform seperti YouTube dan volume interaksi pengguna yang tinggi telah menjadi fokus berbagai penelitian terkait tantangan moderasi konten. Gillespie (2020) menekankan bahwa kebutuhan akan sistem moderasi otomatis menjadi krusial mengingat skala dan kompleksitas platform modern. Analisis terhadap besarnya skala YouTube dan inefisiensi moderasi manual ini menjadi justifikasi utama mengapa penelitian yang berfokus pada otomatisasi filter komentar, seperti yang diusulkan dalam skripsi ini, memiliki urgensi dan relevansi yang tinggi. Lebih lanjut, Dehkhoda & Gunica (2023) mengidentifikasi bahwa komentar negatif dapat secara signifikan merusak pengalaman pengguna dan menghambat diskusi, yang memperkuat argumen perlunya solusi moderasi yang efektif.

Dalam konteks konten spesifik seperti promosi judi *online* (Judol) di Indonesia, penelitian oleh Wulandari & Ramadhany (2023) menunjukkan bahwa platform seperti YouTube kerap dimanfaatkan untuk tujuan tersebut. Hal ini didukung oleh Manaroinsong dkk. (2024) yang menyoroti media sosial sebagai sarana efektif untuk promosi Judol. Temuan ini memperkuat argumen bahwa mekanisme deteksi yang dikembangkan dalam penelitian ini perlu memiliki kemampuan untuk mengenali pola-pola promosi Judol yang spesifik dan seringkali terselubung dalam konteks bahasa dan budaya Indonesia. Studi kriminologi oleh Tumanggor & Yusuf (2025) juga memberikan pemahaman lebih dalam mengenai lanskap permasalahan perjudian *online* di Indonesia, termasuk faktor pendorongnya, yang relevan dalam perancangan strategi deteksi.

Untuk mengatasi keterbatasan dari metode tunggal, pendekatan hibrida yang mengombinasikan beberapa teknik muncul sebagai solusi yang menjanjikan (Yao dkk., 2019). Metode ini memungkinkan sistem untuk memanfaatkan keunggulan dari masing-masing pendekatan, seperti kecepatan deteksi berbasis aturan dan kemampuan generalisasi dari model machine learning. Pendekatan ini sangat relevan dalam analisis komentar daring, di mana sering ditemukan penggunaan bahasa gaul, singkatan, atau frasa yang maknanya sangat bergantung pada konteks budaya lokal. Dengan demikian, integrasi kedua metode dapat meningkatkan akurasi dan mengurangi celah yang mungkin dieksploitasi untuk menghindari deteksi. Oleh karena itu, penelitian ini mengadopsi pendekatan hibrida untuk menggabungkan kekuatan deteksi berbasis aturan yang spesifik untuk konteks lokal dengan analisis toksisitas umum dari API modern.Tabel 2.1 menyajikan perbandingan dari beberapa penelitian terdahulu yang menjadi landasan bagi penelitian ini.

Tabel 2.1 Perbandingan Penelitian Terdahulu.

Peneliti (Tahun)	Judul Penelitian	Metode	Objek/Fokus Penelitian	Hasil/Temuan Utama
Andročec, D. (2020)	*Machine learning methods for toxic comment classification: a systematic review*	Tinjauan Literatur Sistematis	Metode-metode *machine learning* (SVM, CNN, LSTM, dll.) untuk klasifikasi komentar toksik.	Model *deep learning* menunjukkan performa paling efektif, namun sebagian besar penelitian masih berfokus pada dataset Jigsaw dan Bahasa Inggris.
Dehkhoda, S. & Gunica, J. (2023)	*Analyzing Toxicity in YouTube Comments with the Help of Machine Learning*	Analisis Kuantitatif & Kualitatif	Prevalensi dan jenis komentar toksik pada kanal YouTube berbahasa Swedia.	Tingkat toksisitas relatif rendah (0.643%), dengan mayoritas berupa penghinaan personal dan kompetensi, serta menunjukkan bahwa moderasi manual tidak efisien karena volume.
Çakar, B., dkk. (2025)	*Is Reuse All You Need? A Systematic Comparison of Regular Expression Composition Strategies*	Studi Empiris & Perbandingan Sistematis	Membandingkan strategi pembuatan Regex (reuse, sintesis, dan AI generatif) pada tugas-tugas rekayasa perangkat lunak.	penggunaan Regex memiliki akurasi sebanding dengan AI canggih, menyoroti pentingnya basis data Regex yang teruji.
Wulandari, T. F. & Ramadhany, S. W. (2023)	*Ketentuan Hukum Pidana Terhadap Promosi Konten Judi Online*	Yuridis Normatif	Aspek hukum pidana terhadap promosi konten judi *online* (Judol) di Indonesia.	Promosi Judol di platform seperti YouTube secara eksplisit melanggar hukum di Indonesia (UU ITE), namun praktiknya terus meluas.
Yao dkk. (2019)	*Clinical text classification with rule-based features and knowledge-guided convolutional neural networks*	Pendekatan Hibrida (Berbasis Aturan + CNN)	Klasifikasi teks pada catatan klinis medis (bukan komentar *online*).	Membuktikan bahwa pendekatan hibrida yang mengintegrasikan pengetahuan domain (aturan) dengan *deep learning* (CNN) dapat mengungguli metode tunggal.
Hosseini dkk. (2017)	*Deceiving Google's Perspective API Built for Detecting Toxic Comments*	Eksperimen (*Adversarial Attack*)	Kerentanan Google Perspective API terhadap manipulasi teks.	Menunjukkan bahwa API dapat "ditipu" dengan modifikasi teks sederhana (salah ketik, penambahan spasi/tanda baca), yang menurunkan skor toksisitas secara signifikan.

Perbedaan dengan Penelitian yang Dilakukan

Berdasarkan tinjauan pustaka di atas, terlihat bahwa penelitian-penelitian sebelumnya telah memberikan landasan yang kuat dalam berbagai aspek deteksi konten negatif. Andročec (2020) serta Dehkhoda & Gunica (2023) fokus pada analisis toksisitas umum, seringkali dalam konteks bahasa Inggris atau Eropa, dan belum secara spesifik menangani variasi *spam* dalam konteks budaya Indonesia seperti Judol. Di sisi lain, penelitian seperti yang dilakukan oleh Wulandari & Ramadhany (2023) telah membahas Judol dari perspektif hukum, namun tidak menawarkan solusi teknis untuk deteksinya. Sementara itu, penelitian Hosseini dkk. (2017) telah mengidentifikasi kelemahan pada Perspective API, yang menjadi justifikasi untuk tidak hanya mengandalkan API tersebut secara tunggal.
Penelitian ini membedakan diri dan mengisi celah yang ada dengan beberapa poin utama:

1. **Pendekatan Hibrida yang Kontekstual:** Penelitian ini tidak hanya mengandalkan satu metode, tetapi mengombinasikan dua pendekatan yang saling melengkapi. Kekuatan analisis toksisitas umum dari Perspective API digabungkan dengan ketajaman deteksi berbasis aturan (*pattern matching* dan *fuzzy matching*) yang secara spesifik dirancang untuk mengenali pola-pola Judol dan *spam* berbahasa Indonesia. Ini merupakan respons langsung terhadap temuan bahwa API generik memiliki keterbatasan (Hosseini dkk., 2017) dan perlunya penanganan konten spesifik lokal (Wulandari & Ramadhany, 2023).
2. **Fokus pada Implementasi Praktis:** Berbeda dengan banyak penelitian yang berfokus pada evaluasi akademis suatu model, penelitian ini berorientasi pada produk. Hasil akhirnya adalah sebuah **aplikasi *web* fungsional** yang dapat digunakan secara langsung oleh kreator konten YouTube. Ini mengintegrasikan tidak hanya deteksi konten, tetapi juga **tindakan moderasi** (penghapusan komentar) melalui YouTube Data API v3, sebuah aspek implementasi praktis yang seringkali tidak menjadi fokus utama dalam penelitian klasifikasi teks.
3. **Solusi Terintegrasi untuk Kreator Konten:** Penelitian ini tidak hanya bertujuan untuk mengklasifikasikan teks, tetapi untuk menyediakan alat bantu yang komprehensif bagi pengguna akhir (kreator konten). Aplikasi yang dikembangkan menyediakan antarmuka untuk melihat, memfilter, dan mengelola komentar terdeteksi, yang merupakan solusi terintegrasi dari deteksi hingga eksekusi moderasi.

Dengan demikian, kebaruan (*novelty*) dari penelitian ini terletak pada pengembangan sistem hibrida yang dirancang khusus untuk konteks Indonesia (Judol) dan diwujudkan dalam bentuk aplikasi *web* yang praktis dan terintegrasi penuh dengan ekosistem YouTube untuk membantu kreator konten melakukan moderasi secara efisien.

2.2 Dasar Teori

Dasar teori berikut ini memberikan landasan konseptual untuk penelitian yang dilakukan:

2.2.1 YouTube

YouTube adalah platform berbagi video global yang memungkinkan

pengguna mengunggah, menonton, berinteraksi, dan berbagi konten video (Wattenhofer dkk., 2012). Pemahaman akan ekosistem YouTube, termasuk mekanisme interaksi dan volume kontennya, menjadi dasar penting untuk merancang sistem filter komentar yang relevan dan efektif dalam konteks platform ini.

2.2.2 Komentar Online

Komentar *online* adalah teks yang ditulis pengguna sebagai respons terhadap konten digital dan menjadi bentuk utama interaksi di platform seperti YouTube (Wattenhofer dkk., 2012). Penelitian ini berfokus pada identifikasi jenis-jenis komentar berikut:

1. **Komentar Toksik (*Toxic Comments*):** Merujuk pada komentar yang bersifat kasar, tidak sopan, menyerang, atau menghina yang berpotensi membuat pengguna lain merasa tidak nyaman dan enggan melanjutkan diskusi (Dehkhoda & Gunica, 2023). Identifikasi kategori ini penting karena dampak negatifnya yang signifikan terhadap kesehatan dan keberlangsungan komunitas *online*.
2. **Komentar *Spam* (*Spam Comments*):** Mencakup pesan yang tidak relevan dengan konteks konten, berisi promosi komersial yang tidak diinginkan, menyertakan tautan berbahaya, atau merupakan komentar yang dipublikasikan secara berulang-ulang dengan tujuan mengganggu alur diskusi. Deteksi *spam* bertujuan untuk menjaga relevansi dan kualitas interaksi dalam kolom komentar.
3. **Komentar Judi *Online* (Judol):** Merupakan subkategori *spam* yang secara spesifik bertujuan untuk mempromosikan layanan perjudian *online*. Komentar jenis ini sering menggunakan istilah, frasa, dan ajakan khas yang beredar di Indonesia (Wulandari & Ramadhany, 2023; Manaroinsong dkk., 2024). Fokus pada Judol dalam penelitian ini didasari oleh maraknya promosi ilegal tersebut di Indonesia.

2.2.3 Deteksi Konten Negatif

Deteksi konten negatif adalah proses otomatis untuk mengidentifikasi berbagai bentuk konten digital yang dianggap melanggar pedoman komunitas, bersifat ilegal, atau berpotensi berbahaya bagi pengguna. Pendekatan utama yang digunakan dalam penelitian ini meliputi:

1. **Pendekatan Berbasis Aturan/Pola:** Metode ini menggunakan aturan linguistik atau pola tekstual (misalnya, daftar kata kunci spesifik, ekspresi reguler) yang telah didefinisikan secara manual untuk mengidentifikasi konten tertentu (Aubaid & Mishra, 2020). Pendekatan ini dipilih dalam penelitian ini karena kemampuannya untuk menangkap pola-pola Judol yang sangat spesifik dan menggunakan istilah lokal yang mungkin sulit dideteksi secara akurat oleh model statistik umum yang tidak dilatih secara khusus untuk konteks tersebut.
2. **Pendekatan Berbasis *Machine Learning*:** Pendekatan ini melibatkan pelatihan model statistik pada dataset besar berisi teks berlabel untuk mempelajari pola-pola yang membedakan antara konten negatif dan non-negatif, sehingga model dapat mengklasifikasikan teks baru secara otomatis (Andročec, 2020). Meskipun penelitian ini tidak

mengembangkan model ML baru dari awal, pemahaman akan prinsip *machine learning* tetap penting karena Perspective API, yang merupakan salah satu pilar deteksi dalam sistem ini, berbasis pada teknologi tersebut.

3. **Pendekatan Hibrida:** Pendekatan ini mengombinasikan dua atau lebih pendekatan deteksi yang berbeda untuk saling melengkapi, memanfaatkan keunggulan masing-masing, dan menutupi kelemahan yang ada, sehingga dapat meningkatkan kinerja deteksi secara keseluruhan (Yao dkk., 2019; Alyasiri dkk., 2022). Menyadari bahwa setiap pendekatan memiliki kelebihan dan kekurangan masing-masing, Strategi ini dirancang untuk memanfaatkan presisi tinggi dari **pendekatan berbasis aturan** sebagai lapisan pertama untuk menyaring konten dengan pola yang sangat spesifik dan dapat diprediksi, seperti promosi Judol dan *spam*. Kemudian, untuk komentar yang lolos dari saringan pertama, **pendekatan berbasis *machine learning* melalui Perspective API** digunakan sebagai lapisan kedua untuk mengukur tingkat toksisitas yang lebih umum dan kontekstual.

2.2.4 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) atau Pemrosesan Bahasa Alami adalah cabang ilmu komputer dan Artificial Intelligence (AI) yang berfokus pada interaksi antara komputer dan bahasa manusia, dengan tujuan agar komputer dapat memproses, memahami, menginterpretasi, dan bahkan menghasilkan bahasa manusia secara bermakna (Mittal, 2024). Teknik NLP yang relevan dan diterapkan dalam penelitian ini meliputi:

1. **Pencocokan Pola (*Pattern Matching*)**: Penggunaan teknik seperti Ekspresi Reguler (*Regular Expression* atau Regex) untuk menemukan dan mengidentifikasi pola-pola spesifik di dalam teks. Regex adalah sebuah "bahasa" formal untuk mendefinisikan pola pencarian yang sangat kuat dan fleksibel. Daripada mencari kata kunci secara harfiah, Regex memungkinkan pendefinisian aturan kompleks yang dapat menangani berbagai variasi penulisan, menjadikannya sangat berguna untuk pendekatan deteksi berbasis aturan (MDN, 2025).
2. **Pencocokan String Samar (Fuzzy String Matching):** Teknik untuk menemukan string yang mirip atau mendekati string target meskipun tidak ada kecocokan eksak. Pendekatan ini berguna untuk menangani variasi penulisan seperti kesalahan ketik, variasi ejaan informal, atau penggunaan leetspeak yang tidak terstruktur yang mungkin tidak tertangkap oleh aturan regex yang ketat. Penelitian ini memanfaatkan pustaka Fuzzball untuk mengimplementasikan fuzzy matching sebagai pelengkap deteksi berbasis pola Judol/spam.

2.2.5 Perspective API

Perspective API adalah layanan antarmuka pemrograman aplikasi berbasis *cloud* dari Jigsaw (sebuah unit di dalam Google) yang dirancang untuk membantu pengembang dan platform mengelola kualitas percakapan *online* (Hosseini dkk., 2017). API ini menggunakan model *machine learning* untuk menganalisis teks dan memberikan skor probabilitas terkait berbagai atribut yang berpotensi dianggap negatif oleh pembaca. Atribut utama yang dimanfaatkan dalam penelitian ini adalah TOXICITY, namun atribut lain seperti INSULT dan PROFANITY juga dipertimbangkan sebagai penanda tambahan (Nakka, 2025). Perspective API dipilih sebagai salah satu komponen deteksi karena kemampuannya menganalisis toksisitas secara umum pada berbagai jenis teks dan ketersediaannya sebagai layanan eksternal yang dapat diintegrasikan ke dalam aplikasi. Untuk detail lebih lanjut mengenai cara kerja, atribut yang didukung, dan batasan penggunaan API, dokumentasi resmi dari Jigsaw/Google menjadi sumber.

2.2.6 YouTube Data API v3

Youtube Data Api adalah API resmi dari Google yang memungkinkan aplikasi berinteraksi dengan data dan fungsionalitas YouTube secara terprogram. Dalam penelitian ini, API digunakan untuk mengambil daftar video, komentar, dan yang terpenting, melakukan tindakan moderasi seperti penghapusan komentar, dengan otorisasi pengguna. Penggunaan API ini krusial untuk mewujudkan fungsionalitas manajemen komentar dalam aplikasi yang dikembangkan. Detail teknis mengenai penggunaan, batasan kuota, dan proses otorisasi tersedia secara lengkap dalam dokumentasi resmi Google Developers.

2.2.7 Aplikasi Web

Aplikasi *web* merupakan program perangkat lunak yang diakses pengguna melalui peramban *web* standar dan umumnya disimpan pada *server* jarak jauh, berbeda dari aplikasi desktop yang memerlukan instalasi (Oleshchenko, 2021). Dengan arsitektur klien-server, teknologi *front-end* seperti HTML, CSS, dan JavaScript digunakan untuk antarmuka pengguna (Oleshchenko, 2021). Penelitian ini menghasilkan aplikasi *web* untuk moderasi guna memastikan aksesibilitas luas tanpa instalasi khusus.

2.2.8 Application Programming Interface (API)

Dalam pengembangan perangkat lunak modern, aplikasi seringkali perlu berinteraksi dengan perangkat lunak atau layanan lain, dan proses komunikasi ini dimungkinkan oleh sebuah teknologi yang disebut **API (*Application Programming Interface*)**. Menurut MDN Web Docs, API bukanlah *database* atau server itu sendiri, melainkan kode yang mengatur titik akses ke server tersebut, berfungsi sebagai "jembatan" yang memiliki seperangkat aturan dan protokol yang memungkinkan berbagai aplikasi untuk saling berkomunikasi secara terstruktur (MDN, 2024). Penggunaan API memungkinkan pengembang untuk mengintegrasikan layanan eksternal secara efisien dan fokus pada pengembangan logika inti dari aplikasi mereka sendiri.