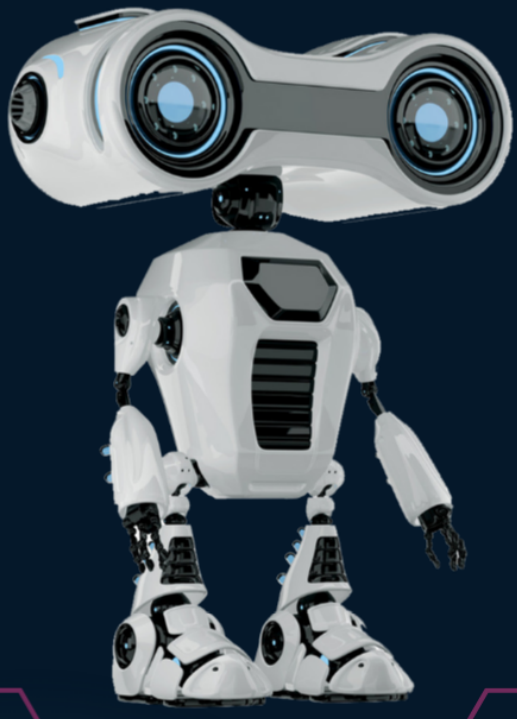
KECERDASAN BUATAN UNTUK MODERASI KONTEN DIGITAL DI MEDIA SOSIAL

DR. ARI MUZAKIR, S.KOM., M. CS.
DR. USMAN EPENDI, M. KOM.
SUYANTO, S.KOM., MM., M. KOM.



KECERDASAN BUATAN UNTUK MODERASI KONTEN DIGITAL DI MEDIA SOSIAL

Oleh:
Dr. Ari Muzakir, S.Kom., M. Cs.
Dr. Usman Ependi, M. Kom.
Suyanto, S.Kom., MM., M. Kom.

Diterbitkan oleh

ASOSIASI DOKTOR
Sistem Informasi
INDONESIA

2025

Kecerdasan Buatan Untuk Moderasi Konten Digital Di Media Sosial

Oleh:
Dr. Ari Muzakir, S.Kom., M. Cs.
Dr. Usman Ependi, M. Kom.
Suyanto, S.Kom., MM., M. Kom.

Uk. 15,5cm x 23cm (vii + 124hlm)

ISBN: 978-634-04-6360-6

Diterbitkan oleh

ASOSIASI DOKTOR Sistem Informasi INDONESIA

Editor: Dr. Aria Hendrawan, S.T., M.Kom

Edisi Desember 2025

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa, atas rahmat dan karunia-Nya buku berjudul *"Kecerdasan Buatan untuk Moderasi Konten Digital di Media Sosial"* ini dapat diselesaikan dengan baik. Buku ini disusun sebagai upaya memberikan pemahaman mengenai pemanfaatan teknologi kecerdasan buatan (AI) dalam menghadapi tantangan moderasi konten di era media sosial yang sarat dengan hoaks, ujaran kebencian, pornografi, perjudian online, dan berbagai bentuk konten berbahaya lainnya.

Secara garis besar, buku ini terdiri dari tiga bagian. Bagian I membahas dasar-dasar moderasi konten digital, termasuk perbedaan moderasi manual dan otomatis. Bagian II menguraikan penerapan AI melalui machine learning, deep learning, NLP, computer vision, speech recognition, serta teknik lanjutan seperti AutoML dan transfer learning. Bagian III menyajikan aplikasi nyata dalam moderasi konten teks, gambar, video, audio, hingga integrasi multimodal. Buku ini diharapkan dapat menjadi rujukan bagi akademisi, peneliti, mahasiswa, praktisi, maupun pembuat kebijakan dalam mengembangkan sistem moderasi konten yang cerdas, efektif, dan beretika.

Pada kesempatan ini, kami menyampaikan ucapan terima kasih yang sebesar-besarnya kepada Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia atas bantuan pendanaan penelitian tahun 2025. Ucapan terima kasih juga kami sampaikan kepada Universitas Bina Darma atas dukungan dan fasilitas yang diberikan selama proses penelitian dan penyusunan panduan ini. Tidak lupa, penghargaan yang tulus kami berikan kepada semua pihak yang telah berkontribusi, baik secara langsung maupun tidak langsung.

Palembang, September 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

== Bagian I ==
DASAR-DASAR MODERASI KONTEN DIGITAL

Bagian ini membahas fondasi utama yang perlu dipahami sebelum memasuki penerapan kecerdasan buatan dalam moderasi konten. Pada bab pertama, dibahas perkembangan media sosial yang sangat pesat dan berbagai tantangan yang muncul dari penyebaran konten digital, mulai dari hoaks, ujaran kebencian, pornografi, hingga judi online. Bab kedua menjelaskan bagaimana proses moderasi dilakukan, baik secara manual maupun otomatis, beserta kelebihan, keterbatasan, dan peralihan menuju pendekatan yang lebih canggih berbasis kecerdasan buatan. Melalui bagian ini, pembaca diharapkan memahami urgensi moderasi konten digital, mengenali jenis-jenis konten berbahaya, serta memperoleh gambaran tentang evolusi metode moderasi dari cara tradisional hingga modern.

BAB I

PENDAHULUAN

Bab ini akan membawa pembaca memahami latar belakang mengapa moderasi konten digital menjadi isu penting di era media sosial. Kita akan melihat bagaimana pesatnya perkembangan platform digital menghadirkan manfaat besar sekaligus tantangan serius, terutama terkait penyebaran konten berbahaya.

Di bagian awal, pembaca akan diajak menelusuri perkembangan media sosial dan kompleksitas konten digital yang terus bertambah, baik dalam bentuk teks, gambar, audio, maupun video. Selanjutnya, pembahasan berfokus pada jenis-jenis konten berbahaya seperti hoaks, ujaran kebencian, pornografi, judi online, hingga penipuan digital yang semakin sering muncul di platform populer.

Akhirnya, bab ini akan menjelaskan mengapa moderasi konten menjadi kebutuhan mendesak bagi keamanan digital dan ketahanan masyarakat. Moderasi bukan hanya persoalan teknis, tetapi juga berkaitan dengan perlindungan generasi muda, stabilitas sosial, serta kepercayaan publik terhadap ruang digital.

Dengan membaca bab ini, pembaca akan mendapatkan gambaran besar tentang masalah utama yang coba diatasi oleh sistem moderasi konten, sebelum melangkah ke pembahasan lebih teknis pada bab-bab selanjutnya.

1.1 Perkembangan Media Sosial dan Tantangan Konten Digital

Media sosial saat ini telah menjadi bagian tak terpisahkan dari kehidupan masyarakat modern. Hampir setiap aktivitas (mulai dari komunikasi sehari-hari, berbagi informasi, hiburan, hingga transaksi ekonomi) dilakukan melalui platform digital seperti Facebook, Instagram, X (Twitter), TikTok, dan YouTube. Di Indonesia, penggunaan media sosial bahkan sangat masif: lebih dari 170 juta orang tercatat aktif menggunakan platform ini setiap bulan (Irawan, Yusufianto, Agustina, 2020). Angka ini menggambarkan betapa

kuatnya peran media sosial dalam membentuk pola pikir, perilaku, bahkan keputusan masyarakat.

Perkembangan media sosial yang pesat ini membawa membawa tantangan baru yang tidak bisa diabaikan. Setiap detik, jutaan konten baru dipublikasikan oleh pengguna dari berbagai belahan dunia. Konten tersebut hadir dalam beragam format: teks singkat dalam bentuk status, gambar yang dilengkapi keterangan, video berdurasi panjang atau pendek, hingga siaran langsung (live streaming) yang berlangsung secara real-time. Dengan volume konten sebesar ini, muncul persoalan krusial: bagaimana membedakan konten positif yang bermanfaat dengan konten berbahaya yang justru merugikan individu maupun masyarakat?

Sifat media sosial yang terbuka semakin memperbesar tantangan tersebut. Siapa pun, tanpa batasan usia atau latar belakang, dapat dengan mudah memproduksi dan menyebarkan informasi. Hal ini membuat media sosial rawan dipenuhi oleh konten yang tidak terverifikasi, menyesatkan, atau bahkan berbahaya. Misalnya, berita palsu (hoaks) bisa menyebar dengan sangat cepat hanya karena dibagikan ulang oleh banyak orang, meskipun sumber awalnya tidak jelas atau sengaja dibuat untuk menyesatkan.

Di sisi lain, karakteristik komunikasi di media sosial cenderung dinamis dan informal. Pengguna seringkali menggunakan singkatan, bahasa gaul, emoji, hingga kode tertentu untuk menyampaikan pesan. Dalam konteks tertentu, hal ini memang membuat komunikasi lebih ekspresif. Namun, dari sisi moderasi konten, gaya komunikasi semacam ini menimbulkan tantangan besar. Sebagai contoh, kata “jp” yang bagi sebagian orang berarti “jackpot” dalam permainan judi online, bisa saja tidak dikenali oleh sistem deteksi sederhana yang hanya mencari kata kunci “judi” atau “betting”.

Tantangan lain datang dari konten yang bersifat multimodal. Sebuah unggahan di Instagram, misalnya, mungkin tidak mengandung kata-kata yang mencurigakan dalam keterangan teksnya. Namun, gambar yang disisipkan ternyata berisi nomor rekening untuk transfer judi atau situs web ilegal. Begitu pula dengan video di TikTok, yang sekilas tampak seperti hiburan biasa, tetapi di dalamnya

terdapat potongan suara yang menyebutkan kode promosi judi online. Kompleksitas ini membuat upaya moderasi manual menjadi sangat sulit dilakukan secara konsisten.

Ilustrasi sederhana bisa digambarkan seperti ini: bayangkan sebuah tim moderator manusia harus memantau jutaan postingan baru setiap jam. Mereka harus membaca teks, memeriksa gambar, mendengarkan audio, bahkan menonton video secara penuh untuk memastikan konten tersebut aman. Tugas ini jelas mustahil dilakukan secara manual, apalagi dengan kecepatan dan volume unggahan yang begitu tinggi. Akibatnya, banyak konten berbahaya lolos dari pengawasan dan akhirnya dikonsumsi oleh publik.

Kondisi ini menegaskan bahwa tantangan utama dalam perkembangan media sosial bukan hanya pada pertumbuhan jumlah pengguna, tetapi juga pada kemampuan untuk menjaga ruang digital tetap aman dan sehat. Moderasi konten menjadi kunci penting, namun metode tradisional tidak lagi memadai. Inilah alasan mengapa kecerdasan buatan (AI) mulai dipandang sebagai solusi strategis untuk membantu proses moderasi konten digital secara lebih efektif dan efisien.

1.2 Jenis Konten Berbahaya di Media Sosial

Media sosial adalah ruang terbuka di mana siapa pun dapat berpartisipasi, sehingga tak hanya konten positif yang beredar, tetapi juga berbagai bentuk konten berbahaya. Konten berbahaya ini memiliki spektrum luas: mulai dari sekadar informasi menyesatkan hingga promosi aktivitas ilegal. Dampaknya pun bisa beragam—mulai dari menurunkan kualitas informasi, memicu konflik sosial, hingga merugikan secara ekonomi dan moral.

Salah satu bentuk konten berbahaya yang paling sering ditemui adalah hoaks dan misinformasi. Contohnya bisa kita lihat pada masa pandemi COVID-19, ketika beredar pesan singkat di WhatsApp dan Facebook yang menyebutkan bahwa bawang putih bisa menyembuhkan virus. Informasi ini tampak sepele, tetapi nyatanya membuat banyak orang terlena dan enggan memeriksakan diri ke fasilitas kesehatan. Dalam kasus lain, berita palsu tentang politik bisa

memicu ketegangan antar pendukung kelompok tertentu, karena disebarkan tanpa klarifikasi yang memadai. Hoaks bekerja seperti api kecil: sekali menyebar, ia bisa dengan cepat membesar dan menimbulkan kerugian nyata.

Selain hoaks, media sosial juga sarat dengan ujaran kebencian. Konten ini biasanya menyerang individu atau kelompok berdasarkan identitas mereka, seperti agama, suku, ras, atau pandangan politik. Misalnya, di sebuah platform sering kita temui komentar bernada kasar terhadap kelompok tertentu hanya karena perbedaan keyakinan. Sekilas mungkin tampak seperti "opini pribadi", tetapi dalam skala besar, ujaran kebencian bisa merusak harmoni sosial. Di beberapa negara, penyebaran ujaran kebencian di media sosial bahkan terbukti berkontribusi pada konflik fisik dan kekerasan di dunia nyata.

Bentuk konten berbahaya lainnya adalah pornografi dan eksploitasi seksual, yang sering menyasar anak-anak dan remaja sebagai target paling rentan. Di TikTok atau Instagram, misalnya, ada akun anonim yang membagikan potongan video dengan konten seksual eksplisit. Meski sering kali platform menutup akun-akun tersebut, mereka dengan cepat bermunculan kembali dengan nama berbeda. Lebih berbahaya lagi, ada kasus eksploitasi seksual anak yang diperdagangkan melalui grup tertutup di aplikasi pesan instan, yang jelas menimbulkan ancaman serius bagi keselamatan generasi muda.

Tidak kalah mengkhawatirkan adalah konten promosi aktivitas ilegal, seperti perdagangan narkoba atau judi online. Judi online, khususnya, kerap menyamarkan diri dalam bentuk iklan terselubung. Contoh konkret: sebuah unggahan di Facebook menampilkan poster berwarna cerah dengan kalimat "Tebak angka keberuntunganmu hari ini, hadiah jutaan menantimu!". Bagi orang awam, mungkin terlihat seperti kuis biasa. Namun, jika diperhatikan lebih detail, terdapat tautan kecil di bagian bawah yang mengarahkan ke situs judi online. Kasus semacam ini menunjukkan bagaimana konten berbahaya bisa tampil dengan wajah yang tampak normal, sehingga sulit dikenali tanpa teknologi pendukung.

Selain empat jenis utama di atas, terdapat pula konten berbahaya lain yang sering muncul di media sosial, seperti penipuan digital (scam). Contoh populer adalah pesan yang menawarkan hadiah undian palsu atau investasi cepat kaya. Banyak pengguna, terutama yang kurang melek digital, menjadi korban jebakan ini. Dampaknya bisa berupa kerugian finansial hingga pencurian data pribadi. Jika kita rangkum, konten berbahaya di media sosial setidaknya dapat dikategorikan ke dalam lima kelompok besar:

1. hoaks dan misinformasi,
2. ujaran kebencian, pornografi dan eksploitasi seksual,
3. promosi aktivitas ilegal seperti narkoba dan judi, serta
4. penipuan digital.

Setiap kategori memiliki karakteristik tersendiri, cara penyamaran berbeda, dan dampak yang sama-sama berbahaya. Oleh karena itu, memahami jenis-jenis konten ini menjadi langkah awal yang penting sebelum masuk ke strategi moderasi dan teknologi yang dapat digunakan untuk mengatasinya.

1.3 Pentingnya Moderasi Konten bagi Keamanan Digital dan Masyarakat

Moderasi konten merupakan salah satu fondasi utama dalam menjaga keamanan digital. Tanpa moderasi, media sosial bisa berubah menjadi "hutan belantara" informasi, di mana siapa pun bebas menanamkan narasi, baik yang positif maupun negatif, tanpa adanya filter. Dalam kondisi demikian, pengguna rentan terpapar berbagai bentuk konten berbahaya yang dapat mengganggu kesehatan mental, merugikan ekonomi, bahkan menimbulkan ancaman terhadap keamanan nasional.

Salah satu fungsi penting moderasi adalah melindungi pengguna dari konten berbahaya. Bayangkan seorang remaja berusia 14 tahun yang sedang menjelajah TikTok. Jika moderasi tidak bekerja, ia bisa dengan mudah menemukan konten pornografi atau iklan judi online yang disamarkan sebagai game. Paparan seperti ini berpotensi merusak perkembangan psikologis anak, mendorong perilaku adiktif, dan dalam jangka panjang memengaruhi pola hidup mereka. Dengan

adanya sistem moderasi, konten semacam ini dapat difilter atau dihapus sebelum sampai ke pengguna yang rentan.

Moderasi juga berperan dalam menjaga ekosistem digital tetap sehat. Media sosial pada dasarnya adalah ruang publik virtual. Jika ruang ini dipenuhi ujaran kebencian, fitnah, atau provokasi, maka atmosfer digital akan berubah menjadi penuh ketegangan. Hal ini dapat memperburuk polarisasi di masyarakat. Sebaliknya, jika moderasi dilakukan dengan baik, platform media sosial dapat menjadi tempat yang aman untuk berdiskusi, belajar, dan membangun jaringan positif.

Selain itu, moderasi konten membantu mendukung penegakan hukum. Konten yang melanggar hukum, seperti promosi narkoba atau perdagangan manusia, seringkali dilakukan secara terselubung di media sosial. Tanpa moderasi yang efektif, aparat penegak hukum kesulitan melacak aktivitas ilegal tersebut. Dengan adanya sistem berbasis kecerdasan buatan, konten mencurigakan bisa terdeteksi sejak dini, kemudian dilaporkan atau ditindaklanjuti oleh pihak berwenang. Contohnya, banyak kasus penipuan digital berhasil diungkap karena laporan dari sistem deteksi otomatis yang digunakan oleh platform.

Pentingnya moderasi juga terlihat dari sisi kepercayaan publik terhadap platform media sosial. Pengguna akan cenderung meninggalkan sebuah platform jika merasa ruang tersebut penuh dengan konten negatif atau berbahaya. Sebaliknya, jika mereka merasa aman, nyaman, dan terlindungi, kepercayaan terhadap platform akan meningkat. Hal ini berdampak langsung pada keberlangsungan bisnis platform itu sendiri. Dengan kata lain, moderasi bukan hanya tanggung jawab sosial, tetapi juga strategi bisnis. Namun, penting untuk diingat bahwa moderasi tidak sekadar soal “menghapus konten berbahaya”. Moderasi juga harus mempertimbangkan aspek kebebasan berekspresi. Ada kalanya sebuah konten berada di wilayah abu-abu—apakah ia sekadar opini atau sudah termasuk ujaran kebencian? Apakah sebuah video berfungsi sebagai satire atau benar-benar mengandung hoaks? Menyeimbangkan antara kebebasan berekspresi dan kebutuhan

untuk melindungi masyarakat adalah tantangan besar dalam moderasi konten.

Dengan demikian, moderasi konten digital bukan sekadar isu teknis, melainkan isu sosial, hukum, dan etika yang kompleks. Ia menyangkut perlindungan generasi muda, menjaga kohesi sosial, memperkuat kepercayaan terhadap ruang digital, dan mendukung penegakan hukum. Oleh sebab itu, moderasi konten menjadi pilar penting dalam membangun ekosistem media sosial yang sehat, aman, dan berkelanjutan.

BAB II
MODERASI KONTEN: MANUAL VS OTOMATIS

Setelah memahami pentingnya moderasi konten di Bab 1, pada bab ini pembaca akan diajak mendalami berbagai pendekatan moderasi yang digunakan di media sosial dari masa awal hingga saat ini. Bab ini menjelaskan bagaimana metode moderasi berkembang dari yang paling sederhana hingga menuju pemanfaatan kecerdasan buatan.

Pembahasan dimulai dengan moderasi manual, di mana manusia berperan langsung sebagai peninjau konten. Bagian ini akan menunjukkan kelebihan manusia dalam memahami konteks komunikasi, sekaligus keterbatasannya dalam skala besar dan risiko psikologis yang ditimbulkan.

Selanjutnya, bab ini membahas moderasi berbasis aturan (rule-based) yang menggunakan daftar kata kunci atau pola tertentu untuk mendeteksi konten berbahaya. Metode ini lebih cepat dan konsisten dibanding manual, namun kaku dan sering gagal menangkap konteks sebenarnya.

Akhirnya, bab ini menyoroti peralihan menuju moderasi otomatis berbasis kecerdasan buatan (AI). Pendekatan ini menawarkan fleksibilitas, adaptasi terhadap pola baru, serta kemampuan menangani konten multimodal (teks, gambar, audio, video). Namun, pembahasan juga akan menyinggung tantangan etika dan bias data yang masih menyertainya. Dengan mempelajari bab ini, pembaca akan memahami evolusi moderasi konten digital: dari sistem yang sepenuhnya bergantung pada manusia, beralih ke rule-based, hingga ke era modern dengan dukungan AI yang lebih adaptif.

2.1 Moderasi Manual: Kelebihan dan Keterbatasan

Pada masa awal perkembangan media sosial, moderasi konten hampir sepenuhnya dilakukan secara manual oleh manusia. Moderator bertugas meninjau setiap laporan pengguna atau memantau langsung unggahan yang dianggap bermasalah. Proses ini menyerupai “penjaga gerbang” yang memastikan konten berbahaya tidak lolos ke ruang publik digital.

Kelebihan utama moderasi manual terletak pada kemampuan memahami konteks secara mendalam. Seorang moderator bisa menilai apakah sebuah postingan adalah sekadar humor satir, karya

seni, atau benar-benar bermaksud menyebarkan ujaran kebencian. Misalnya, jika seseorang menulis komentar "Dasar kamu alien!", moderator dapat mengenali ini sebagai candaan antar teman, bukan bentuk penghinaan. Fleksibilitas semacam ini sulit dicapai oleh sistem otomatis berbasis kata kunci yang cenderung kaku dan tidak mampu menangkap nuansa komunikasi manusia. Namun, moderasi manual menghadapi tantangan besar dalam hal skala dan kecepatan (Gongane et al., 2022). Bayangkan sebuah platform global seperti Facebook atau TikTok yang menerima miliaran unggahan setiap harinya. Mustahil bagi moderator manusia untuk memeriksa seluruh konten tersebut satu per satu. Bahkan jika ribuan moderator direkrut, jumlah konten yang harus disaring akan tetap jauh lebih besar dibandingkan kemampuan manusia. Akibatnya, banyak konten berbahaya terlewat atau telat ditindak.

Selain persoalan skala, pekerjaan ini juga memiliki dampak psikologis yang berat. Moderator sering kali harus berhadapan dengan konten ekstrem: video kekerasan, pornografi anak, ujaran kebencian, hingga propaganda terorisme. Paparan berulang terhadap konten semacam ini dapat menyebabkan stres, gangguan tidur, bahkan trauma psikologis yang serius. Beberapa penelitian mencatat bahwa sebagian moderator mengalami gejala *post-traumatic stress disorder* (PTSD) akibat pekerjaannya. Hal ini menunjukkan bahwa moderasi manual bukan hanya tidak efisien, tetapi juga berbahaya bagi kesehatan mental mereka yang menjalankannya.

Ilustrasi sederhana bisa digambarkan seperti ini: bayangkan seorang moderator harus meninjau 500–700 postingan per hari. Dalam satu jam, ia mungkin menemukan konten lucu atau informatif, tetapi di jam berikutnya ia bisa saja melihat gambar kekerasan yang mengganggu. Siklus seperti ini berlangsung setiap hari, membuat pekerjaan tersebut menjadi salah satu profesi yang paling rentan terhadap kelelahan mental di era digital. Dengan keterbatasan tersebut, jelas bahwa moderasi manual hanya efektif pada skala kecil atau sebagai pendamping sistem lain. Pada platform besar, ketergantungan penuh pada moderator manusia tidak realistis. Oleh karena itu, dibutuhkan metode lain yang dapat bekerja lebih cepat dan

efisien, sekaligus mengurangi beban psikologis yang harus ditanggung manusia. Transisi inilah yang kemudian melahirkan sistem moderasi berbasis aturan dan, pada tahap berikutnya, berbasis kecerdasan buatan.

Studi Kasus Moderasi Manual di Industri Media Sosial

Salah satu contoh paling dikenal datang dari Facebook. Pada 2018, terungkap bahwa perusahaan ini mempekerjakan ribuan moderator eksternal di berbagai negara, termasuk di Filipina, India, dan Irlandia. Para moderator ini bertugas meninjau laporan pengguna terkait konten berbahaya. Setiap harinya, mereka harus menyaring ratusan hingga ribuan postingan yang berisi kekerasan, ujaran kebencian, pornografi, bahkan konten bunuh diri.

Meskipun pekerjaan ini krusial, banyak moderator melaporkan kondisi kerja yang sangat berat. Sebuah laporan investigatif menyebutkan bahwa moderator hanya memiliki waktu beberapa detik untuk menentukan apakah sebuah konten harus dihapus, ditandai, atau dibiarkan. Tekanan untuk mengambil keputusan cepat seringkali menimbulkan kesalahan. Lebih jauh lagi, banyak moderator mengalami trauma psikologis setelah terpapar konten ekstrem secara terus-menerus. Beberapa di antaranya mengalami mimpi buruk, kecemasan, dan gejala PTSD (*post-traumatic stress disorder*).

Kasus serupa juga muncul pada TikTok. Pada 2020, beberapa mantan moderator TikTok di Amerika Serikat menggugat perusahaan karena tidak memberikan perlindungan kesehatan mental yang memadai. Mereka mengaku harus menonton ratusan video mengerikan setiap hari, mulai dari kekerasan fisik hingga eksploitasi seksual. Untuk mengurangi risiko, sebagian perusahaan kini memberikan fasilitas konseling psikologis dan membatasi durasi kerja moderator. Namun, solusi ini hanya mengurangi dampak, bukan menghilangkan masalah mendasarnya: pekerjaan tersebut memang sangat berisiko bagi mental manusia.

Dari dua studi kasus ini, jelas terlihat bahwa moderasi manual, meski unggul dalam memahami konteks, memiliki keterbatasan fundamental. Ia tidak bisa mengimbangi skala konten yang terus

bertambah, dan ia membebankan risiko besar terhadap kesehatan mental pekerja. Situasi ini semakin menegaskan bahwa moderasi manual hanya bisa berperan sebagai lapisan tambahan untuk mengatasi kasus-kasus kompleks yang membutuhkan penilaian manusia, sementara kebutuhan utama harus ditangani dengan sistem otomatis berbasis teknologi.

2.2 Moderasi Berbasis Aturan (Rule-Based)

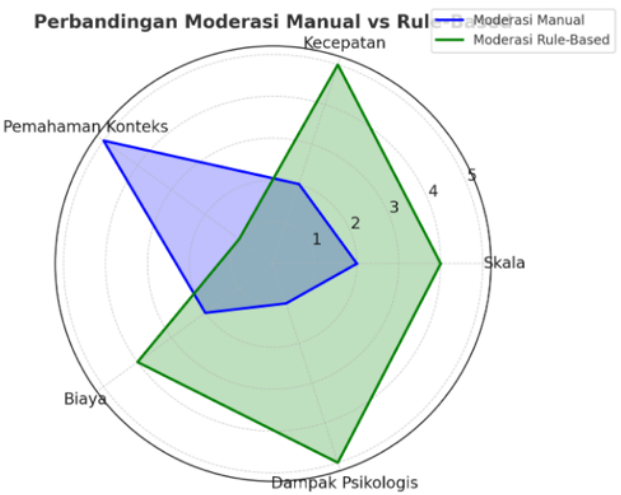
Metode rule-based adalah langkah awal menuju moderasi otomatis yang lebih canggih (Chandrasekaran & Mago, 2021). Ia membantu mengurangi beban moderator manusia, mempercepat proses deteksi, dan memberikan konsistensi dalam penegakan aturan. Namun, keterbatasan dalam memahami konteks, bahasa gaul, serta variasi budaya lokal membuat metode ini sering tidak memadai. Oleh karena itu, rule-based biasanya digunakan sebagai lapisan pertama penyaringan, yang kemudian dilanjutkan dengan sistem berbasis kecerdasan buatan (AI) untuk menghasilkan moderasi yang lebih akurat dan adaptif.

Ketika skala media sosial semakin besar dan moderasi manual terbukti tidak mampu mengimbangi jumlah konten yang beredar, platform digital mulai mengembangkan pendekatan rule-based atau berbasis aturan. Metode ini menggunakan seperangkat aturan yang ditentukan sebelumnya untuk menyaring konten. Biasanya, aturan ini berbentuk daftar kata kunci, pola teks, atau metadata tertentu yang dianggap berbahaya.

Cara kerjanya sederhana: jika sebuah unggahan mengandung kata “judi”, “bom”, atau “pornografi”, sistem akan menandainya secara otomatis. Begitu pula jika ada gambar dengan tanda-tanda eksplisit tertentu (misalnya, deteksi warna kulit pada area tubuh tertentu), sistem rule-based bisa langsung memblokir unggahan tersebut. Dengan metode ini, platform dapat menyaring jutaan unggahan dalam waktu singkat, tanpa harus menunggu campur tangan manusia terlebih dahulu.

Kelebihan dari moderasi rule-based adalah kecepatan dan konsistensi. Sistem dapat bekerja 24 jam nonstop, dengan hasil yang

relatif seragam. Misalnya, jika aturan menyatakan bahwa semua komentar dengan kata "narkoba" harus dihapus, maka sistem akan menegakkan aturan itu tanpa ragu, berbeda dengan moderator manusia yang mungkin masih mempertimbangkan konteks. Rule-based juga lebih murah dibandingkan menambah ribuan tenaga moderator manual. Namun, kelemahan mendasar dari metode ini adalah ketidakmampuannya memahami konteks. Sebagai contoh, jika ada artikel ilmiah yang membahas dampak buruk judi online dengan kalimat: *"Penelitian ini menganalisis maraknya praktik judi online di Indonesia"*, sistem rule-based bisa saja langsung menandainya sebagai konten berbahaya hanya karena ada kata "judi online". Sebaliknya, konten promosi judi yang menggunakan bahasa gaul seperti "scatter", "maxwin", atau "jp" bisa lolos karena tidak ada di daftar kata kunci.



Gambar 2. 1 Perbandingan Moderasi Manual dan Rule-Based di Media Sosial

Ilustrasi ini menunjukkan bahwa meskipun rule-based unggul dalam efisiensi, ia tidak bisa sepenuhnya menggantikan moderasi manual karena kelemahannya dalam memahami konteks. Oleh sebab itu, banyak platform media sosial mengombinasikan keduanya sebelum beralih ke pendekatan berbasis kecerdasan buatan (AI) yang lebih adaptif. Gambar tersebut memperlihatkan perbandingan antara dua metode moderasi konten: manual (biru) dan rule-based (hijau),

berdasarkan lima aspek utama, yaitu: skala, kecepatan, pemahaman konteks, biaya, dan dampak psikologis.

1. Skala & Kecepatan: Moderasi manual lemah dalam skala dan kecepatan karena bergantung pada kapasitas manusia. Sebaliknya, rule-based dapat memproses jutaan konten dengan cepat dan konsisten.
2. Pemahaman Konteks: Moderasi manual unggul karena manusia mampu memahami nuansa bahasa, satire, atau parodi. Rule-based cenderung kaku, hanya mengandalkan kata kunci sehingga mudah salah dalam menilai konteks.
3. Biaya: Moderasi manual membutuhkan banyak tenaga kerja, sehingga mahal. Rule-based lebih murah karena sistem otomatis bisa menggantikan sebagian besar pekerjaan manusia.
4. Dampak Psikologis: Moderator manual berisiko tinggi mengalami kelelahan mental atau trauma akibat paparan konten berbahaya. Rule-based hampir tidak memiliki risiko psikologis karena dijalankan oleh mesin.

Tabel 2. 1 Perbandingan Moderasi Manual dan Rule-Based

Aspek	**Moderasi Manual**	**Moderasi Rule-Based**
Skala	Terbatas, hanya bisa menangani sedikit konten per hari	Dapat menangani jutaan konten sekaligus
Kecepatan	Lambat, tergantung jumlah moderator manusia	Sangat cepat, bekerja otomatis 24/7
Pemahaman Konteks	Tinggi, mampu membedakan satire, humor, atau ujaran kebencian	Rendah, kaku, hanya berdasarkan kata kunci
Biaya	Tinggi, butuh banyak tenaga kerja dan pelatihan	Lebih murah setelah sistem dibangun

Dampak Psikologis	Tinggi, berisiko trauma karena paparan konten ekstrem	Tidak ada, karena dikerjakan oleh mesin
Fleksibilitas	Tinggi, manusia dapat menyesuaikan penilaian	Rendah, sulit menyesuaikan dengan dinamika bahasa baru

Studi Kasus Moderasi Rule-Based

Salah satu kasus nyata terjadi pada platform YouTube. Pada awalnya, YouTube menggunakan sistem rule-based sederhana untuk menandai video yang mengandung kata-kata sensitif di judul atau deskripsi. Akibatnya, banyak kreator konten edukatif yang membahas isu-isu penting, seperti "pencegahan narkoba" atau "penanggulangan pornografi anak", justru mengalami demonetisasi atau penghapusan video. Hal ini memicu protes karena sistem dianggap tidak adil dan merugikan pembuat konten yang sebenarnya ingin memberikan edukasi positif.

Kasus lain muncul di Twitter (kini X). Sistem rule-based mereka sering kali salah membaca konteks bahasa gaul. Misalnya, kata "bakar" yang digunakan anak muda dalam konteks "membakar semangat" pernah ditandai sebagai konten berbahaya karena sistem mengenalinya sebagai indikasi kekerasan. Hal ini memperlihatkan bagaimana sistem rule-based, meskipun cepat, bisa menghasilkan kesalahan klasifikasi (false positive) maupun kebocoran konten (false negative).

2.3 Peralihan Menuju Moderasi Otomatis Berbasis AI

Seiring meningkatnya jumlah konten digital dan semakin canggihnya modus penyamaran konten berbahaya, jelas bahwa moderasi manual maupun rule-based tidak lagi memadai. Manual tidak mampu mengikuti skala besar dan menimbulkan dampak psikologis, sementara rule-based terlalu kaku dan mudah dilewati dengan penggunaan bahasa gaul atau kode. Inilah yang mendorong lahirnya moderasi berbasis kecerdasan buatan (AI) sebagai solusi yang lebih adaptif dan efisien.

Kelebihan utama AI terletak pada kemampuannya belajar dari data. Berbeda dengan sistem rule-based yang hanya mengikuti aturan statis, AI dapat dilatih dengan dataset yang berisi berbagai contoh konten berbahaya. Dengan pembelajaran ini, sistem mampu mengenali pola baru yang sebelumnya tidak ditentukan oleh manusia. Misalnya, jika istilah "jp" dan "scatter" sering muncul dalam konteks promosi judi online, model AI dapat mengaitkannya dengan aktivitas perjudian meskipun kata "judi" tidak pernah disebutkan secara eksplisit.

AI juga menawarkan fleksibilitas dalam menghadapi konten multimodal. Sebuah unggahan di media sosial tidak lagi terbatas pada teks saja, tetapi bisa berupa kombinasi gambar, video, dan audio. Dalam hal ini, teknologi Computer Vision memungkinkan deteksi gambar eksplisit, Optical Character Recognition (OCR) bisa membaca teks dalam gambar, sementara Speech Recognition dapat mengonversi suara dalam video menjadi teks untuk dianalisis. Dengan pendekatan ini, sistem moderasi AI bisa menilai sebuah unggahan dari berbagai sisi secara bersamaan, sesuatu yang tidak mungkin dilakukan oleh sistem rule-based sederhana.

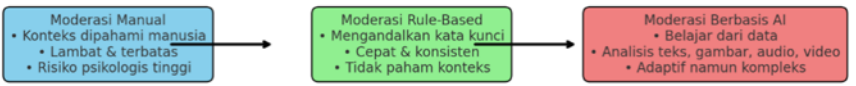
Contoh nyata peralihan ke moderasi berbasis AI bisa dilihat pada YouTube. Platform ini menggunakan model deep learning untuk mendeteksi video berisi ujaran kebencian atau kekerasan. Alih-alih hanya mencari kata kunci di deskripsi video, AI menganalisis isi audio, ekspresi wajah, bahkan objek di dalam video untuk menentukan apakah konten tersebut melanggar aturan. Begitu juga dengan Facebook, yang menggunakan AI untuk mendeteksi konten berbahaya dalam lebih dari 40 bahasa, sehingga lebih adaptif terhadap keragaman budaya dan bahasa lokal.

Meski demikian, moderasi berbasis AI bukan berarti tanpa tantangan. Pertama, AI sangat bergantung pada data latih. Jika dataset tidak cukup representatif, hasil deteksi bisa bias. Misalnya, model yang hanya dilatih dengan data bahasa Inggris mungkin tidak akurat ketika digunakan untuk mendeteksi ujaran kebencian dalam bahasa Indonesia. Kedua, AI masih kesulitan memahami sarkasme, satire, atau konteks budaya tertentu. Sebuah komentar yang bernada humor

bisa saja ditandai sebagai ujaran kebencian, atau sebaliknya, ujaran kebencian yang disamarkan dengan humor bisa lolos dari filter.

Tantangan lainnya adalah terkait etika dan transparansi. Pengguna seringkali tidak tahu bagaimana AI membuat keputusan. Ketika sebuah unggahan dihapus secara otomatis, muncul pertanyaan: siapa yang menentukan standar "berbahaya" atau "tidak pantas"? Apakah sistem bisa salah dan merugikan kebebasan berekspresi? Inilah alasan mengapa AI harus dilengkapi dengan mekanisme pelaporan, banding, dan pengawasan manusia sebagai penyeimbang.

Dengan segala kelebihan dan tantangannya, jelas bahwa peralihan menuju moderasi berbasis AI bukan sekadar tren, tetapi sebuah kebutuhan strategis. AI mampu mengisi kekosongan yang tidak bisa ditangani oleh manusia maupun sistem rule-based. Namun, keberhasilan penerapannya bergantung pada kualitas data, pemahaman konteks lokal, serta keseimbangan antara keamanan digital dan kebebasan berekspresi.



Gambar 2. 2 Evolusi Moderasi Konten Digital: Manual → Rule-Based → AI

Gambar diagram ini menggambarkan evolusi pendekatan moderasi konten di media sosial dari masa awal hingga saat ini:

1. Moderasi Manual. Dilakukan sepenuhnya oleh manusia. Kelebihannya adalah kemampuan memahami konteks dengan baik, misalnya membedakan antara satire, humor, atau ujaran kebencian. Namun, kelemahannya terletak pada kecepatan dan skala yang terbatas, serta risiko psikologis tinggi bagi moderator karena harus berhadapan langsung dengan konten berbahaya.
2. Moderasi Rule-Based. Menggunakan aturan sederhana berbasis kata kunci atau pola teks. Kelebihannya adalah cepat, konsisten, dan mampu memproses jutaan unggahan dalam waktu singkat. Akan tetapi, sistem ini kaku dan tidak mampu memahami konteks, sehingga sering salah mendeteksi (false positive/false negative).

3. Moderasi Berbasis AI. Tahap terbaru yang memanfaatkan kecerdasan buatan. Sistem AI dapat belajar dari data, mengenali pola baru, serta menganalisis berbagai jenis konten (teks, gambar, audio, video). Moderasi AI lebih adaptif dan efisien, meski tetap menghadapi tantangan terkait bias data, keterbatasan pemahaman konteks budaya, serta isu etika dan transparansi.

== Bagian II ==
KECERDASAN BUATAN UNTUK MODERASI KONTEN

Bagian ini berfokus pada bagaimana teknologi kecerdasan buatan (Artificial Intelligence/AI) menjadi inti dari sistem moderasi konten digital modern. Bab ketiga memperkenalkan dasar-dasar AI yang relevan, termasuk perbedaan antara machine learning dan deep learning, serta penerapan Natural Language Processing (NLP), computer vision, dan speech recognition. Bab keempat mengulas berbagai algoritma populer yang digunakan, mulai dari metode klasik seperti SVM, Naive Bayes, hingga deep learning dengan CNN, RNN, LSTM, dan Transformer. Selain itu, dibahas pula konsep AutoML, transfer learning, dan model bahasa modern seperti BERT dan IndoBERT. Bab kelima menekankan pentingnya data sebagai bahan utama moderasi konten, mencakup sumber data, proses anotasi, tantangan linguistik, serta aspek etika penggunaan data. Dengan mempelajari bagian ini, pembaca akan memperoleh pemahaman teknis sekaligus praktis mengenai bagaimana AI bekerja untuk menganalisis teks, gambar, video, dan audio, serta bagaimana keterbatasan dapat diatasi melalui inovasi teknologi.

BAB III
DASAR-DASAR KECERDASAN BUATAN DALAM MODERASI KONTEN

Setelah memahami evolusi moderasi konten dari manual hingga rule-based di bab sebelumnya, pada bab ini pembaca akan diperkenalkan pada dasar-dasar kecerdasan buatan (AI) yang menjadi fondasi utama moderasi konten digital modern.

Pembahasan dimulai dengan perbedaan Machine Learning (ML) dan Deep Learning (DL), dua pendekatan utama dalam AI. Di sini akan dijelaskan bagaimana ML bekerja dengan fitur sederhana, sementara DL mampu belajar langsung dari data kompleks seperti teks panjang, gambar, maupun video.

Selanjutnya, bab ini memperkenalkan tiga cabang AI yang sangat relevan dengan moderasi konten:

- Natural Language Processing (NLP): teknologi untuk memahami bahasa manusia, termasuk slang, singkatan, dan ujaran kebencian.
- Computer Vision (CV): teknologi untuk "melihat" gambar dan video, misalnya mendeteksi pornografi atau membaca teks tersembunyi dengan OCR.
- Speech Recognition: teknologi untuk mengubah ucapan menjadi teks, sehingga konten audio atau video bisa dianalisis lebih lanjut.

Dengan mempelajari bab ini, pembaca akan memperoleh landasan konseptual tentang bagaimana AI bekerja dalam mengenali konten digital berbahaya. Pemahaman dasar ini penting sebagai pijakan sebelum masuk ke Bab 4, yang akan membahas algoritma populer dan teknik lanjutan yang digunakan dalam sistem moderasi berbasis AI.

3.1 Machine Learning vs Deep Learning

Kecerdasan buatan (*Artificial Intelligence*/AI) adalah payung besar yang mencakup berbagai pendekatan agar mesin mampu menirukan proses berpikir manusia (Russell & Norvig, 2002). Dari berbagai pendekatan tersebut, Machine Learning (ML) dan Deep Learning (DL) adalah dua metode yang paling banyak digunakan dalam moderasi konten digital. Keduanya sama-sama berbasis pada

data, tetapi berbeda dalam cara kerja, kompleksitas, serta kemampuannya memahami pola.

Machine Learning (ML) bekerja dengan cara melatih model berdasarkan *fitur* tertentu dari data. Fitur adalah representasi sederhana dari data yang dianggap penting. Misalnya, untuk mendeteksi ujaran kebencian, sistem ML bisa dilatih dengan contoh kalimat positif dan negatif. Dari data tersebut, algoritma belajar mengenali pola seperti frekuensi kata kasar, kombinasi kata tertentu, atau panjang kalimat. Model klasik ML yang sering digunakan antara lain *Naive Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest*.

Kelebihan ML adalah sederhana, cepat dilatih, dan bekerja baik untuk dataset yang relatif kecil. Namun, kelemahannya adalah ketergantungan pada pemilihan fitur yang tepat. Jika fitur tidak dirancang dengan baik, akurasi sistem akan rendah. Misalnya, jika hanya menggunakan daftar kata kasar, model ML bisa gagal mengenali ujaran kebencian yang disamarkan dengan ejaan kreatif seperti "j*di" atau "ba**g".

Deep Learning (DL), di sisi lain, bekerja dengan arsitektur *neural networks* yang meniru cara kerja otak manusia. Teknologi ini mampu mengekstrak pola kompleks secara otomatis, tanpa perlu programmer menentukan fitur secara manual. Misalnya, sebuah model *Convolutional Neural Network (CNN)* dapat mengenali gambar pornografi dengan mempelajari jutaan contoh gambar, tanpa programmer harus menetapkan aturan "gambar dengan warna kulit dominan" atau bentuk tertentu. Model lain seperti *Recurrent Neural Network (RNN)* atau *Transformer* bahkan bisa memahami struktur bahasa dalam teks panjang, termasuk nuansa konteks yang sulit ditangkap oleh ML tradisional.

Kelebihan DL adalah kemampuannya belajar langsung dari data mentah (teks, gambar, audio, video), sehingga hasilnya lebih akurat dan adaptif. Namun, kelemahannya adalah kebutuhan akan data dalam jumlah besar dan komputasi yang kuat. DL tidak akan optimal jika hanya dilatih dengan data yang sedikit. Perbedaan keduanya bisa dilihat dari contoh sederhana. Jika kita ingin mendeteksi komentar bernada kebencian di Twitter:

- Algoritma ML mungkin membutuhkan daftar kata kasar sebagai fitur, lalu menghitung seberapa sering kata-kata itu muncul.
- Model DL seperti *Transformer* dapat langsung memahami makna kalimat secara keseluruhan, bahkan ketika kata-kata tersebut disamarkan dengan ejaan gaul atau digunakan dalam konteks bercanda.

Dengan kata lain, ML adalah seperti "aturan berbasis fitur" yang harus ditentukan manusia, sedangkan DL adalah "pembelajar mandiri" yang bisa menemukan pola lebih kompleks dari data itu sendiri. Keduanya sama-sama penting dalam moderasi konten, namun DL semakin dominan seiring meningkatnya volume dan keragaman data di media sosial.

Tabel 3. 1 Perbandingan Machine Learning (ML) dan Deep Learning (DL) (Muzakir, 2024)

Aspek	**Machine Learning (ML)**	**Deep Learning (DL)**
Kompleksitas	Relatif sederhana, berbasis fitur buatan manusia	Kompleks, menggunakan jaringan saraf tiruan (neural networks) dengan banyak lapisan
Kebutuhan Data	Bekerja cukup baik dengan dataset kecil hingga menengah	Membutuhkan dataset besar agar performa optimal
Ekstraksi Fitur	Manual, fitur harus ditentukan oleh peneliti	Otomatis, model belajar langsung dari data mentah
Kecepatan Training	Cepat, cocok untuk komputer dengan sumber daya terbatas	Lambat, membutuhkan GPU/TPU dengan daya komputasi tinggi
Akurasi	Cukup baik untuk pola sederhana	Sangat tinggi untuk pola kompleks (teks panjang, gambar, audio, video)

Konteks & Adaptasi	Terbatas, sulit menangkap nuansa bahasa gaul/sarkasme	Lebih adaptif, dapat memahami konteks dan variasi bahasa
Contoh Algoritma	Naive Bayes, SVM, Random Forest	CNN, RNN, LSTM, Transformer
Kelebihan	Mudah diterapkan, tidak butuh data besar	Sangat akurat, mampu memproses data multimodal
Kelemahan	Tergantung fitur buatan manusia, rentan salah klasifikasi	Butuh data besar & komputasi mahal, sulit dijelaskan (black box)

📌 **Ringkasan:**

- ML cocok digunakan pada kasus dengan data terbatas dan kebutuhan cepat.
- DL unggul untuk kasus kompleks dengan data besar seperti moderasi teks gaul, gambar eksplisit, atau analisis video.

3.2 Natural Language Processing (NLP)

Sebagian besar konten di media sosial berbentuk teks seperti status, komentar, caption, atau cuitan singkat. Untuk memproses dan memahaminya, kecerdasan buatan menggunakan cabang ilmu yang disebut Natural Language Processing (NLP). NLP adalah teknologi yang memungkinkan komputer membaca, memahami, dan menganalisis bahasa manusia (Lopez & Kalita, 2017).

Di dunia nyata, teks di media sosial tidak selalu rapi seperti di buku atau artikel ilmiah. Pengguna sering menulis dengan singkatan, bahasa gaul, campuran bahasa (code-switching), atau bahkan salah ejaan. Misalnya, kalimat:

"Lu kalah trus, mending coba scatter biar JP gampang 😎🔥"

Bagi orang awam, ini hanyalah obrolan biasa. Namun, sistem NLP yang terlatih bisa mengenali bahwa kata "scatter" dan "JP" sering digunakan dalam konteks promosi judi online. Tanpa NLP, kalimat ini mungkin lolos karena tidak mengandung kata "judi" secara eksplisit. NLP bekerja melalui beberapa tahap utama:

1. Preprocessing Teks. Teks yang berantakan perlu dibersihkan terlebih dahulu. Misalnya, menghapus tanda baca berlebih, mengubah kata ke bentuk dasar (*stemming/lemmatization*), atau menerjemahkan emoji menjadi kata yang relevan.
2. Representasi Kata. NLP modern menggunakan teknik *word embeddings* seperti Word2Vec, GloVe, atau FastText untuk merepresentasikan kata dalam bentuk vektor numerik. Dengan ini, komputer bisa memahami bahwa kata "marah" lebih dekat maknanya dengan "kesal" dibandingkan dengan "senang".
3. Pemahaman Konteks. Model berbasis Transformer seperti BERT atau IndoBERT dapat memahami makna kata dalam konteks kalimat. Misalnya, kata "bakar" dalam kalimat *"Ayo bakar semangatmu!"* tidak sama artinya dengan "bakar" dalam kalimat *"Kita bakar rumah mereka"*.

Dalam moderasi konten digital, NLP digunakan untuk berbagai keperluan, antara lain:

- Deteksi ujaran kebencian (hate speech detection). Sistem NLP bisa menganalisis komentar dan menandai ujaran diskriminatif atau kasar.
- Deteksi hoaks dan misinformasi. Model NLP dilatih dengan data berita benar dan palsu untuk mengidentifikasi pola penyebaran informasi menyesatkan.
- Deteksi iklan terselubung. Dengan bantuan OCR (untuk teks dalam gambar), NLP menganalisis kata-kata promosi ilegal, misalnya terkait judi atau narkoba.

Contoh Kasus Nyata:

Facebook menggunakan NLP dalam lebih dari 40 bahasa untuk mendeteksi ujaran kebencian. Sementara itu, di Indonesia, peneliti banyak memanfaatkan IndoBERT—model bahasa berbasis Transformer yang dilatih khusus pada teks bahasa Indonesia—untuk mendeteksi ujaran kebencian, hoaks, hingga spam iklan. Dengan NLP, proses moderasi teks yang tadinya memerlukan ribuan moderator manusia bisa dilakukan secara otomatis dengan kecepatan dan akurasi yang jauh lebih tinggi.

Dengan NLP, mesin tidak lagi sekadar membaca kata per kata, tetapi mampu memahami makna dalam konteks komunikasi manusia yang kompleks. Inilah sebabnya NLP menjadi salah satu komponen inti dalam sistem moderasi konten digital.

3.3 Computer Vision (CV)

Selain teks, konten berbahaya di media sosial juga banyak muncul dalam bentuk gambar dan video. Untuk mengenali dan memahami konten visual ini, kecerdasan buatan menggunakan cabang ilmu bernama Computer Vision (CV). CV memungkinkan mesin untuk "melihat" dunia digital seperti mata manusia—menganalisis objek, pola, bahkan teks yang tersembunyi dalam gambar atau video.

Di media sosial, gambar dan video sering digunakan untuk menyembunyikan pesan berbahaya. Misalnya, sebuah poster berwarna cerah yang sekilas terlihat seperti iklan game, ternyata menyisipkan link judi online di pojok kecil. Atau sebuah meme lucu yang beredar di Instagram ternyata berisi ujaran kebencian dalam bentuk teks yang ditulis dengan font khusus. Tanpa CV, konten semacam ini bisa lolos dari moderasi karena tidak ada kata mencurigakan dalam deskripsi teks.

CV bekerja melalui beberapa komponen utama:

1. Deteksi Objek. AI dapat mengenali apakah sebuah gambar mengandung objek tertentu, seperti senjata, narkoba, atau bagian tubuh eksplisit. Algoritma populer untuk tugas ini adalah Convolutional Neural Network (CNN) yang mampu menganalisis pola visual dengan sangat baik.

2. Optical Character Recognition (OCR). Teknologi ini memungkinkan mesin membaca teks yang ada di dalam gambar atau video. Misalnya, jika sebuah poster berisi tulisan "slot gacor maxwin malam ini", OCR akan mengekstrak teks tersebut, lalu sistem NLP dapat menganalisis maknanya lebih lanjut.
3. Analisis Video. Video diproses frame demi frame. Dengan CV, sistem bisa mendeteksi apakah video mengandung adegan kekerasan, pornografi, atau simbol-simbol terlarang. Teknologi ini juga bisa dipadukan dengan *speech recognition* untuk menganalisis suara dalam video.

Contoh Kasus Nyata:

- YouTube menggunakan CV untuk mendeteksi video berisi kekerasan atau konten seksual eksplisit. Dalam banyak kasus, video bisa langsung ditandai hanya beberapa menit setelah diunggah, bahkan sebelum dilaporkan oleh pengguna.
- Instagram memanfaatkan CV untuk mendeteksi gambar yang berpotensi berbahaya, termasuk iklan judi online yang disamarkan sebagai konten hiburan.
- Penelitian di Indonesia juga banyak menggunakan kombinasi OCR + NLP untuk mendeteksi teks iklan judi yang disisipkan di gambar atau video pendek TikTok.

Kelebihan CV adalah kemampuannya menangani konten visual yang semakin mendominasi media sosial. Namun, ada juga tantangan yang perlu dihadapi. Pertama, CV kadang menghasilkan kesalahan ketika gambar dibuat dengan teknik manipulasi (misalnya deepfake). Kedua, CV bisa bias jika data latihnya tidak beragam. Misalnya, model yang hanya dilatih dengan dataset Barat mungkin gagal mengenali konten berbahaya dengan konteks lokal di Asia Tenggara.

Dengan Computer Vision, sistem moderasi konten tidak hanya bisa membaca kata-kata, tetapi juga melihat gambar dan video yang mengandung konten berbahaya. Teknologi ini melengkapi NLP,

sehingga moderasi bisa dilakukan lebih menyeluruh pada berbagai jenis konten di media sosial.

3.4 Speech Recognition

Selain teks dan gambar, konten berbahaya di media sosial juga banyak ditemukan dalam bentuk audio dan video. Dengan semakin populernya platform seperti TikTok, Instagram Reels, dan podcast, audio menjadi medium penting dalam penyebaran informasi, termasuk konten bermasalah. Untuk menganalisisnya, kecerdasan buatan menggunakan teknologi Speech Recognition atau pengenalan suara.

Speech Recognition adalah teknologi yang memungkinkan komputer mengubah ucapan menjadi teks, sehingga isi percakapan atau narasi dalam audio bisa dipahami dan dianalisis lebih lanjut dengan teknik NLP. Dengan cara ini, sistem dapat mendeteksi kata-kata atau frasa berbahaya yang disampaikan secara lisan, bukan hanya dalam bentuk tulisan. Contoh sederhana: dalam sebuah video live streaming, seorang streamer berkata,

"Klik link di bio buat dapet scatter maxwin malam ini!"

Jika sistem hanya menganalisis caption video, kemungkinan besar tidak akan menemukan kata-kata terkait judi. Namun, dengan Speech Recognition, ucapan streamer bisa ditranskripsi menjadi teks, lalu dianalisis menggunakan NLP. Sistem kemudian dapat menandai konten tersebut sebagai promosi judi online.

Speech Recognition dalam Moderasi Konten bisa digunakan untuk berbagai tujuan:

1. Deteksi ujaran kebencian berbasis suara. Misalnya, dalam sebuah rekaman debat atau diskusi, AI dapat mengenali kalimat diskriminatif yang diucapkan oleh salah satu pembicara.
2. Deteksi promosi aktivitas ilegal. Banyak pelaku judi online menyamarkan iklannya dalam bentuk ucapan singkat di video. Speech Recognition membantu mengidentifikasi pola ini.

3. Kombinasi multimodal. Speech Recognition biasanya dipadukan dengan CV (untuk visual) dan NLP (untuk teks). Dengan integrasi ini, sistem bisa menganalisis teks, gambar, dan suara secara bersamaan dalam satu unggahan.

Contoh Kasus Nyata:

- YouTube menggunakan teknologi Speech-to-Text untuk membuat subtitle otomatis pada video. Selain mempermudah aksesibilitas, data transkrip ini juga digunakan untuk mendeteksi konten berbahaya yang diucapkan pembuat video.
- TikTok mulai mengembangkan sistem deteksi otomatis untuk audio, terutama dalam memfilter musik atau narasi yang mengandung kata-kata sensitif.
- Di Indonesia, penelitian terkini memanfaatkan Speech Recognition untuk mengekstrak promosi judi yang diselipkan dalam video pendek, kemudian mengombinasikannya dengan NLP untuk klasifikasi.

Tantangan utama Speech Recognition dalam moderasi adalah keragaman bahasa, aksen, dan gaya bicara. Misalnya, seorang pengguna mungkin mencampur bahasa Indonesia dan Inggris dalam satu kalimat (code-switching), atau menggunakan slang yang tidak umum. Kesalahan transkripsi bisa membuat sistem gagal mendeteksi konten berbahaya. Selain itu, kualitas audio yang buruk, adanya musik latar, atau suara berisik juga bisa memengaruhi akurasi.

Dengan Speech Recognition, sistem moderasi menjadi lebih lengkap dan menyeluruh. Teknologi ini memastikan bahwa konten berbahaya yang disisipkan dalam bentuk audio juga bisa dideteksi, bukan hanya teks dan gambar.

BAB IV
ALGORITMA DAN TEKNIK POPULER

Setelah memahami dasar-dasar kecerdasan buatan dalam moderasi konten pada bab sebelumnya, bab ini akan membawa pembaca lebih dalam ke berbagai algoritma dan teknik populer yang menjadi tulang punggung sistem moderasi digital berbasis AI.

Pembahasan diawali dengan algoritma klasifikasi teks tradisional, seperti *Naive Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest*. Algoritma ini banyak digunakan pada tahap awal penelitian moderasi konten karena sederhana dan efektif untuk dataset yang terbatas.

Selanjutnya, fokus akan beralih ke teknik deep learning, termasuk *Convolutional Neural Networks (CNN)*, *Recurrent Neural Networks (RNN)*, *Long Short-Term Memory (LSTM)*, hingga arsitektur modern berbasis *Transformer*. Model-model ini telah terbukti unggul dalam memahami bahasa alami, menganalisis gambar, serta menangani data multimodal.

Bab ini juga akan membahas AutoML, sebuah pendekatan yang memudahkan pengguna dalam memilih model terbaik secara otomatis, serta transfer learning dan model bahasa besar (seperti *BERT* dan *IndoBERT*) yang memungkinkan pemanfaatan pengetahuan dari data berskala global untuk kebutuhan lokal.

Dengan mempelajari bab ini, pembaca akan memahami alat dan teknik praktis yang bisa diterapkan dalam membangun sistem moderasi konten digital. Pengetahuan ini menjadi jembatan antara teori AI dan penerapannya di dunia nyata, yang akan semakin diperdalam pada Bab 5 mengenai data serta Bab 6–8 mengenai aplikasi moderasi pada teks, gambar, audio, dan multimodal.

4.1 Klasifikasi Teks (SVM, Random Forest, Naive Bayes)

Pada tahap awal moderasi konten digital, banyak penelitian menggunakan algoritma klasifikasi tradisional untuk mendeteksi konten berbahaya. Algoritma ini bekerja dengan menganalisis teks berdasarkan fitur sederhana, seperti frekuensi kata, kombinasi kata (*n-gram*), atau panjang kalimat. Meskipun tidak sekuat *deep learning*,

metode ini cukup efektif untuk data terbatas dan menjadi fondasi awal perkembangan moderasi konten otomatis.

4.1.1 Naive Bayes

Naive Bayes adalah algoritma berbasis probabilitas yang mengasumsikan setiap fitur independen satu sama lain. Kesederhanaannya membuat algoritma ini sangat cepat dilatih dan diimplementasikan. Naive Bayes (NB) sendiri merupakan keluarga klasifier probabilistik yang didasarkan pada Aturan Bayes dengan asumsi *conditional independence* (fitur dianggap independen satu sama lain diberikan kelas) (Murugesan & Kaliyamurthie, 2023). Untuk dokumen *d* dengan fitur kata *x1,x2,…,xn* dan kelas *C*, kita gunakan:

$$P(C \mid d) = \frac{P(C)\,P(d \mid C)}{P(d)} \quad \Rightarrow \quad P(C \mid d) \propto P(C)\,P(d \mid C)$$

Dengan asumsi *naive* (independensi kondisional), untuk Multinomial NB:

$$P(d \mid C) = \prod_{i=1}^{V} P(w_i \mid C)^{\text{count}(w_i \text{ dalam } d)}$$

di mana $P(w_i \mid C)$ dihitung dari frekuensi kata pada kelas *C*. Untuk menghindari probabilitas nol, dipakai *Laplace smoothing* (α = 1 biasa dipakai).

Langkah-langkah Algoritma (Training & Prediksi)

- ✓ **Training (Multi-NB untuk teks):**
  1. Preprocessing teks: lowercase, tokenisasi, (opsional) stopword removal, stemming/lemmatization, opsi: TF vs counts.
  2. Bangun vocabulary *V* dari semua kata pelatihan.
  3. Untuk tiap kelas *C*:
     - Hitung jumlah dokumen N_C .
     - Hitung total token kata dalam kelas: $T_C=\sum_w \text{count}(_w,C)$.
     - Hitung frekuensi tiap kata dalam kelas: count $(_w,C)$.
  4. Hitung prior kelas: $P(C)=NC/N_{total}$
  5. Hitung likelihood (dengan Laplace α):

$$P(w \mid C) = \frac{\text{count}(w, C) + \alpha}{T_C + \alpha \cdot |V|}$$

6. Simpan *P(C)* dan semua $P(_w \mid C)$.

✓ **Prediksi (untuk dokumen baru *d*):**

1. Preprocess dokumen menjadi token dan hitung count per kata.
2. Untuk tiap kelas *C* hitung skor log untuk menghindari underflow:

$$\text{score}(C) = \log P(C) + \sum_{w \in d} \text{count}(w, d) \cdot \log P(w \mid C)$$

3. Pilih kelas dengan score tertinggi. (Jika ingin probabilitas normalkan exponent dari log-score.)

Contoh Perhitungan Lengkap (langkah demi langkah):
Kita pakai contoh kecil supaya semua langkah bisa dihitung manual dan dibaca jelas.
Vocabulary (V) = 8 kata:
`daftar, menang, jp, slot, gratis, diskon, maxwin, fashion`

✓ **Dokumen Pelatihan (label *G* = Gambling, *N* = Non-Gambling)**

- Kelas G (3 dokumen):
  G1: daftar menang jp
  G2: slot jp maxwin
  G3: menang jp
- Kelas N (3 dokumen):
  N1: gratis diskon fashion
  N2: diskon fashion gratis
  N3: fashion gratis

Hitung jumlah dokumen total = 6 → priors:

- $P(G) = 3/6 = 0.5$
- $P(N) = 3/6 = 0.5$

Hitung frekuensi kata per kelas:
Kelas G — jumlah token dan frekuensi

- daftar: 1
- menang: 2

- jp: 3
- slot: 1
- maxwin: 1
- gratis: 0
- diskon: 0
- fashion: 0

  Total token TG=1+2+3+1+1+0+0+0 = 8

Kelas N — frekuensi

- gratis: 3
- diskon: 2
- fashion: 3
- daftar, menang, jp, slot, maxwin: 0

  Total token TN=3+2+3+0+0+0+0+0=8

Pakai Laplace smoothing α = 1.

Hitung $P(_w|C)$ (tabel):

Untuk kelas G (rumus (count+1)/16:

- daftar: (1+1)/16 = 2/16 = 0.125
- menang: (2+1)/16 = 3/16 = 0.1875
- jp: (3+1)/16 = 4/16 = 0.25
- slot: (1+1)/16 = 2/16 = 0.125
- maxwin: (1+1)/16 = 2/16 = 0.125
- gratis: (0+1)/16 = 1/16 = 0.0625
- diskon: 1/16 = 0.0625
- fashion: 1/16 = 0.0625

Untuk kelas N:

- gratis: (3+1)/16 = 4/16 = 0.25
- diskon: (2+1)/16 = 3/16 = 0.1875
- fashion: (3+1)/16 = 4/16 = 0.25
- daftar, menang, jp, slot, maxwin: (0+1)/16 = 1/16 = 0.0625 masing-masing

✓ **Prediksi: klasifikasikan dokumen baru**

Dokumen baru d^* = “jp maxwin daftar”

Kita hitung un-normalized posterior untuk tiap kelas:

$$\text{score\_unnorm}(C) = P(C) \cdot \prod_{w \in d^*} P(w \mid C)$$

Untuk kelas G:

- $P(G) = 0.5$
- Likelihood $P(d^* \mid G) = P(\text{jp} \mid G) \times P(\text{maxwin} \mid G) \times P(\text{daftar} \mid G)$ $= 0.25 \times 0.125 \times 0.125.$

Perkalian langkah-per-langkah:

1. 0.25 × 0.125=0.031250.25 (Penjelasan: 0.25=1/4, 025 = 1/8.(1/4)×(1/8)=1/32=0.03125
2. 0.03125 × 0.125 = 0.00390625. (Karena 0.03125 = 1/32, dikali 1/8 → 1/256=0.00390625.)

Jadi P(d*|G)=0.00390625.

Kemudian dikali prior:
score_unnorm(G)=0.5×0.00390625=0.001953125
(0.5 × 0.00390625 = 0.001953125 — juga sama dengan 1/512.)

Untuk kelas N:

- $P(N) = 0.5$
- $P(d^* \mid N) = P(\text{jp} \mid N) \times P(\text{maxwin} \mid N) \times P(\text{daftar} \mid N)$ $= 0.0625 \times 0.0625 \times 0.0625.$ (Karena jp, maxwin, daftar masing-masing 1/16 = 0.0625.)

Perkalian:

1. 0.0625×0.0625=0.00390625 (Karena 1/16×1/16=1/256.)
2. 0.00390625×0.0625=0.000244140625. (Karena 1/256 × 1/16 =1/4096.)

Jadi P(d*|N)=0.000244140625.
Dikalikan prior:
score_unnorm(N)=0.5×0.000244140625=0.0001220703125.
(Itu sama dengan 1/8192.)

Normalisasi (menghitung probabilitas posterior):
Jumlah unnormalized:
S=0.001953125+0.0001220703125=0.0020751953125.
Probabilitas posterior:

- P(G|d*)=0.001953125/0.0020751953125. Kita ubah ke bentuk pecahan supaya mudah:
  Kita sebelumnya lihat 0.001953125 = 1/512 dan 0.0001220703125=1/8192

Maka sum S = 1/512 + 1/8192 = (16/8192) + (1/8192) = 17/8192

Jadi P(G|d*) = (1/512)/(17/8192) = (1/512)×(8192/17) = 16/17 ≈ 0.9411764705882353.

- P(N|d*) = 1/17 ≈ 0.058823529411764705.

Kesimpulan prediksi: model mengklasifikasikan dokumen sebagai Gambling (G) dengan probabilitas ≈ 94.1%.
Catatan: hasil eksak 16/17 muncul karena angka-angka pada contoh dibuat agar menjadi pecahan yang rapi — ini memudahkan demonstrasi matematika.

Kenapa kita pakai Laplace smoothing?

Tanpa smoothing, jika satu kata tidak pernah muncul dalam kelas tertentu maka P(w|C)=0 dan seluruh produk P(d|C) menjadi nol → kelas tidak akan pernah dipilih. Laplace α=1 menambahkan 1 ke semua hitungan sehingga mencegah nol. (α bisa diset lain mis. 0.1, tergantung dataset.)

Tabel 4. 1 Kelebihan dan Keterbatasan Naive Bayes dalam Moderasi Konten

Aspek	**Penjelasan**
Kelebihan	- Sederhana, cepat, efektif pada dataset kecil hingga menengah - Membutuhkan sedikit sumber daya (CPU/memori), mudah diimplementasikan - Sering menjadi baseline yang kuat untuk klasifikasi teks
Keterbatasan	- Asumsi independensi fitur sering tidak realistis, meskipun praktis masih cukup baik - Kurang optimal pada tugas yang memerlukan pemahaman konteks semantik yang kompleks (lebih cocok dengan DL) - Performa sensitif terhadap

	pemrosesan fitur (tokenisasi, stopwords, n-grams, TF vs counts)

Pseudocode ringkas (Multinomial Naive Bayes)

```
Training(docs, labels, alpha=1):
  V = build_vocabulary(docs)
  for each class C:
    N_C = number of docs with label C
    T_C = total token count in docs of class C
    for w in V:
      count_w_C = total occurrences of w in docs of class C
      P_w_given_C = (count_w_C + alpha) / (T_C + alpha * |V|)
    P_C = N_C / N_total
  return model {P_C, P_w_given_C for all C,w}

Predict(doc, model):
  tokens = preprocess(doc)
  for each class C:
    logscore[C] = log(P_C)
    for each token w in tokens:
      if w in V:
        logscore[C] += count(w in doc) * log(P_w_given_C)
      else:
     # Unknown word: either skip or use small probability (alpha/(T_C + alpha*|V|))
  return argmax_C logscore[C]
```

Rekomendasi praktis untuk aplikasi moderasi teks

- Gunakan Multinomial NB untuk fitur count kata; Bernoulli NB bila menggunakan fitur kehadiran/ketidakhadiran (bag-of-words biner).
- Cobalah k-gram (unigram + bigram) untuk menangkap frasa (mis. "judi online", "slot gacor").
- Terapkan *Laplace smoothing* ($\alpha \geq 1$) untuk mencegah nol.
- Jika dataset besar & bahasa gaul/kontekstual banyak, pertimbangkan upgrade ke Transformer (BERT/IndoBERT); gunakan NB sebagai baseline cepat.
- Evaluasi: gunakan metrik akurasi, precision, recall, F1; analisis kesalahan untuk menemukan kata/phrases yang sering bikin salah klasifikasi.

4.1.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma klasifikasi yang berusaha mencari garis pemisah (hyperplane) terbaik untuk memisahkan data dari dua kelas. Hyperplane ini dipilih sedemikian rupa sehingga margin (jarak antara hyperplane dengan titik terdekat dari tiap kelas, yang disebut support vectors) adalah maksimum.
Intuisinya: SVM tidak hanya mencari garis pemisah sembarang, tetapi garis yang "paling aman" sehingga jika ada data baru masuk, kemungkinan besar tetap bisa dipisahkan dengan benar.

- Untuk data 2 dimensi, hyperplane adalah sebuah garis.
- Untuk data 3 dimensi, hyperplane adalah sebuah bidang.
- Untuk dimensi lebih tinggi (misalnya data teks yang memiliki ribuan fitur), hyperplane adalah bidang berdimensi tinggi.

Langkah-langkah Algoritma SVM (Linear Case)

1. Representasi Data. Setiap teks (dokumen) diubah menjadi vektor fitur, misalnya dengan metode bag-of-words atau TF-IDF.
2. Menentukan Hyperplane. Algoritma mencoba mencari persamaan:
   $$w \cdot x + b = 0$$
   di mana:
   - w = bobot (vektor normal hyperplane)
   - x = vektor data
   - b = bias
3. Kondisi Margin Maksimum. Untuk setiap data:
   - Jika label $y_i = +1$, maka $w \cdot x_i + b \geq +1$
   - Jika label $y_i = -1$, maka $w \cdot x_i + b \leq -1$

   Tujuan optimisasi: memaksimalkan margin $2/||w||$.
4. Support Vectors. Titik-titik yang berada tepat di tepi margin disebut support vectors. Titik-titik inilah yang "menentukan" garis pemisah.
5. Kernel Trick (Non-Linear Case). Jika data tidak bisa dipisahkan secara linear, SVM menggunakan kernel (misalnya polynomial kernel, RBF kernel) untuk memproyeksikan data ke ruang berdimensi lebih tinggi agar bisa dipisahkan.

Contoh Ilustrasi Sederhana
Misalkan kita ingin melatih SVM untuk mendeteksi teks sederhana dengan hanya dua fitur:

- x1: jumlah kata “slot” dalam teks
- x2: jumlah kata “gratis” dalam teks

Tabel 4. 2 Contoh Dataset

Teks	**x1 (slot)**	**x2 (gratis)**	**Label**
“main slot jp”	1	0	+1 (judi)
“slot gratis maxwin”	1	1	+1 (judi)
“diskon gratis fashion”	0	1	-1 (non-judi)
“promo fashion murah gratis”	0	1	-1 (non-judi)

Visualisasi (di bidang 2D):

- Titik kelas +1 (judi): (1,0) dan (1,1)
- Titik kelas -1 (non-judi): (0,1) dan (0,1)

SVM akan mencari garis pemisah terbaik, misalnya mendekati:
x1=0.5
yang memisahkan teks dengan kata “slot” (kelas judi) dan teks tanpa kata “slot” (kelas non-judi).

Tabel 4. 3 Kelebihan dan Keterbatasan Support Vector Machine (SVM) dalam Moderasi Konten

Aspek	**Penjelasan**
Kelebihan	- Efektif untuk data berdimensi tinggi seperti teks (ribuan fitur kata) - Tidak mudah overfitting karena hanya berfokus pada support vectors - Bekerja baik pada dataset kecil hingga menengah dengan margin kelas yang jelas
Keterbatasan	- Proses pelatihan lambat pada dataset yang sangat besar - Sulit diinterpretasi, terutama jika menggunakan kernel yang kompleks - Kurang cocok untuk data dengan noise tinggi atau kelas yang saling tumpang tindih (overlap)

Dengan contoh ini, kita bisa lihat bahwa SVM sangat cocok untuk deteksi ujaran kebencian atau iklan terselubung berbasis teks dengan fitur sederhana. Namun, untuk data sosial media yang lebih kompleks (bahasa gaul, multimodal), biasanya SVM digantikan oleh deep learning.

4.1.3 Random Forest

Random Forest adalah algoritma klasifikasi berbasis ensemble learning yang menggabungkan banyak Decision Tree untuk menghasilkan prediksi yang lebih stabil dan akurat.

- Decision Tree tunggal membuat keputusan dengan membagi data berdasarkan aturan sederhana (misalnya, apakah teks mengandung kata "slot", apakah panjang kalimat lebih dari 10 kata, dsb.). Namun, pohon tunggal sering terlalu spesifik (overfitting) terhadap data latih.
- Random Forest memperbaiki hal ini dengan membangun banyak Decision Tree, masing-masing dilatih dengan subset acak dari data dan fitur. Hasil prediksi akhir ditentukan berdasarkan voting mayoritas dari semua pohon.

Langkah-langkah Kerja Random Forest

1. Ambil sampel data acak (*bootstrap sample*) dari dataset.
2. Bangun Decision Tree dari sampel tersebut dengan memilih subset fitur acak pada setiap percabangan.
3. Ulangi proses hingga terbentuk ratusan atau ribuan pohon.
4. Untuk prediksi dokumen baru, setiap pohon memberikan klasifikasi, lalu hasil akhir diputuskan berdasarkan voting terbanyak.

Contoh Ilustrasi

Misalkan kita ingin mendeteksi konten judi online. Fitur sederhana yang digunakan:

- Apakah teks mengandung kata "slot"?
- Apakah teks mengandung kata "maxwin"?
- Apakah teks mengandung kata "gratis"?
- Pohon 1 memutuskan berdasarkan kata "slot" → jika ada, prediksi *judi*.
- Pohon 2 memutuskan berdasarkan kata "maxwin" → jika ada, prediksi *judi*.

- Pohon 3 memutuskan berdasarkan kata “gratis” → jika ada, prediksi *non-judi*.

Jika dokumen berisi kata “slot” dan “gratis”, maka:

- Pohon 1 → *judi*
- Pohon 2 → *tidak judi* (tidak ada “maxwin”)
- Pohon 3 → *non-judi*

Hasil voting: 2 *non-judi* vs 1 *judi* → klasifikasi akhir: *non-judi*.
Dengan banyak pohon, Random Forest cenderung lebih stabil dan akurat daripada satu Decision Tree.

Tabel 4. 4 Kelebihan dan Keterbatasan Random Forest dalam Moderasi Konten

Aspek	**Penjelasan**
Kelebihan	- Lebih stabil daripada satu Decision Tree karena hasil didasarkan pada kombinasi banyak pohon - Tahan terhadap *overfitting*, terutama jika jumlah pohon cukup banyak - Dapat menangani fitur dalam jumlah besar dengan baik - Memberikan estimasi pentingnya fitur (*feature importance*), misalnya menunjukkan kata yang paling berpengaruh dalam klasifikasi ujaran kebencian
Keterbatasan	- Lebih lambat dibanding Decision Tree tunggal, terutama jika jumlah pohon sangat banyak - Kurang interpretatif, sulit dipahami karena melibatkan ratusan pohon - Masih kalah dengan *deep learning* untuk teks yang kompleks dan sarat konteks semantik

4.2 Deep Learning (CNN, RNN, LSTM, Transformer)

Seiring meningkatnya kompleksitas data di media sosial—baik teks yang panjang, bahasa gaul yang terus berubah, gambar yang sarat simbol, video dengan konten eksplisit, hingga audio berisi ujaran kebencian—algoritma klasik seperti Naive Bayes, SVM, dan Random Forest mulai menunjukkan keterbatasannya. Algoritma klasik cenderung hanya mampu mengenali pola sederhana, misalnya keberadaan kata kunci tertentu, tetapi sering gagal memahami

konteks, makna tersirat, atau kombinasi multimodal (teks, gambar, audio sekaligus).

Untuk mengatasi tantangan ini, muncullah pendekatan Deep Learning (DL). DL bekerja dengan jaringan saraf tiruan berlapis (neural networks) yang meniru cara kerja otak manusia. Setiap lapisan jaringan mampu mengekstrak pola dari data: lapisan awal menangkap fitur sederhana (seperti huruf atau kata), lapisan menengah mengenali pola lebih besar (frasa, kalimat, atau bentuk visual), dan lapisan lebih dalam memahami representasi yang kompleks (konteks kalimat, makna ujaran, bahkan emosi dalam suara).

Keunggulan utama DL adalah kemampuannya belajar langsung dari data mentah, tanpa perlu fitur buatan manusia. Jika algoritma klasik membutuhkan "daftar kata kasar" atau "aturan manual" untuk mendeteksi ujaran kebencian, model DL dapat menemukan pola sendiri dengan menganalisis ribuan atau jutaan contoh. Hal ini membuat DL jauh lebih fleksibel dan adaptif terhadap variasi bahasa serta konten kreatif di media sosial. Dalam konteks moderasi konten, DL digunakan pada berbagai skenario:

- Teks: memahami nuansa ujaran kebencian, ironi, atau sarkasme.
- Gambar: mendeteksi konten pornografi, kekerasan, atau simbol terlarang.
- Video: menganalisis adegan visual sekaligus teks/ucapan di dalamnya.
- Audio: mengenali ujaran toksik atau promosi ilegal dalam siaran langsung maupun rekaman.

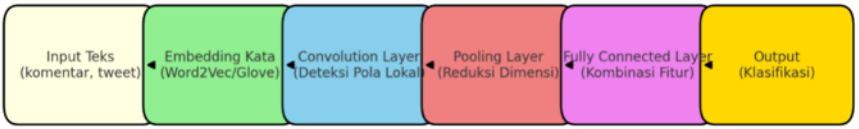
Perkembangan DL juga memunculkan berbagai arsitektur khusus sesuai jenis data yang diproses. Misalnya, CNN unggul dalam mendeteksi pola visual dan teks pendek, RNN memahami urutan kata dalam kalimat, LSTM mengingat konteks panjang, sementara Transformer menjadi standar baru yang mampu memahami teks kompleks dengan efisien dan akurat.

Dengan kemampuannya yang luas, DL menjadi tulang punggung sistem moderasi konten modern. Namun, kehebatan ini datang

dengan konsekuensi: kebutuhan akan dataset besar, komputasi tinggi, serta risiko bias jika data latih tidak beragam. Oleh karena itu, memahami berbagai jenis model DL beserta kekuatan dan keterbatasannya adalah kunci untuk membangun sistem moderasi konten yang efektif dan bertanggung jawab.

4.2.1 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) adalah jenis jaringan saraf tiruan yang awalnya dirancang untuk pengolahan gambar. CNN bekerja dengan cara mengekstrak pola lokal melalui operasi konvolusi, di mana filter (*kernel*) bergerak di atas data untuk mendeteksi fitur penting seperti tepi, bentuk, atau tekstur. Setelah beberapa lapisan konvolusi, jaringan dapat mengenali pola yang semakin kompleks. Berikut diagram alur CNN untuk moderasi teks: mulai dari **input teks → embedding → convolution → pooling → fully connected → output klasifikasi.**



Gambar 4. 1 Alur Convolutional Neural Network (CNN) untuk Moderasi Teks

Diagram ini menggambarkan proses kerja CNN dalam menganalisis teks. Teks masukan seperti komentar atau tweet terlebih dahulu diubah menjadi representasi numerik melalui embedding kata. Lapisan konvolusi kemudian mendeteksi pola lokal penting, misalnya frasa bernada kasar atau kata kunci promosi judi. Hasilnya diringkas dalam lapisan pooling, lalu digabungkan pada fully connected layer untuk menghasilkan prediksi akhir. Output sistem berupa klasifikasi, misalnya *konten berbahaya* atau *konten normal*.

Walaupun populer di visi komputer, CNN juga terbukti efektif untuk teks. Dalam konteks teks, filter konvolusi bisa digunakan untuk mendeteksi frasa atau kombinasi kata yang khas, misalnya frasa “slot gacor” atau “maxwin hari ini” dalam iklan judi online.

Arsitektur CNN (sederhana)

1. **Input Layer.** Data berupa teks (dikonversi menjadi representasi numerik, misalnya *word embeddings*) atau gambar (piksel).
2. **Convolution Layer.** Filter mendeteksi pola lokal, misalnya kombinasi kata atau pola visual.
3. **Pooling Layer**. Mereduksi dimensi data dengan mengambil nilai maksimum (*max pooling*) agar hanya pola penting yang tersisa.
4. **Fully Connected Layer**. Menggabungkan semua fitur yang diekstrak untuk menghasilkan prediksi akhir.
5. **Output Layer**. Menghasilkan kelas akhir, misalnya *judi* vs *non-judi*, atau *konten eksplisit* vs *normal*.

Contoh Kasus CNN dalam Moderasi Konten

- **Moderasi Teks:** CNN dilatih untuk mendeteksi ujaran kebencian dalam komentar Facebook. Filter dapat mengenali pola kata kasar walaupun muncul dalam posisi berbeda di kalimat.
- **Moderasi Gambar:** CNN mendeteksi poster berisi promosi judi online dengan pola visual khas (warna mencolok, ikon slot, teks "scatter", dsb.).
- **Moderasi Video:** CNN menganalisis frame demi frame untuk menemukan adegan eksplisit atau simbol terlarang.

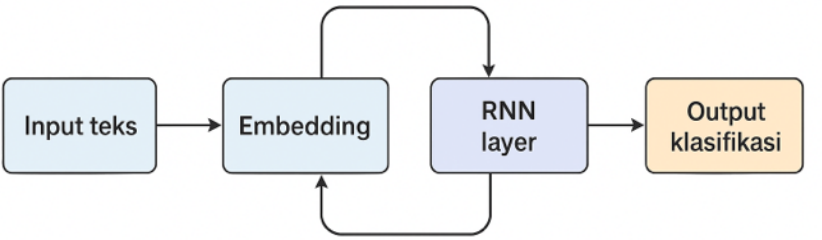
Tabel 4. 5 Kelebihan dan Keterbatasan CNN dalam Moderasi Konten

Aspek	**Penjelasan**
Kelebihan	- Mampu mendeteksi pola lokal dengan baik - Cepat dilatih dibanding model sekuensial - Cocok untuk data teks pendek maupun gambar - Didukung oleh banyak framework modern
Keterbatasan	- Sulit memahami konteks panjang dalam teks - Membutuhkan dataset besar untuk stabilitas - Rentan gagal jika pola disamarkan (misalnya ejaan gaul)

4.2.2 Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) adalah jenis jaringan saraf tiruan yang dirancang untuk memproses data sekuensial atau berurutan, seperti teks, suara, sinyal waktu, atau video. Berbeda dengan jaringan saraf konvensional seperti CNN, RNN memiliki mekanisme *feedback loop* yang memungkinkan informasi dari langkah sebelumnya disimpan dan digunakan pada langkah berikutnya. Mekanisme ini membuat RNN sangat kuat untuk menangkap dependensi temporal atau konteks dalam data berurutan.
Dalam moderasi konten, RNN digunakan untuk memahami makna teks yang bergantung pada urutan kata, sehingga mampu membedakan nuansa kalimat. Misalnya, kalimat **"Saya tidak suka judi"** memiliki makna berbeda dari **"Saya suka judi"**, dan konteks ini penting untuk klasifikasi konten.

Diagram Alur RNN untuk Moderasi Teks
Diagram alur RNN secara umum terdiri dari tahapan:
Input teks → Word Embedding → RNN Layer → Fully Connected Layer → Output klasifikasi



Gambar 4. 2 Alur Recurrent Neural Network (RNN) untuk Moderasi Teks

Diagram ini menggambarkan proses kerja RNN dalam moderasi teks. Prosesnya dapat dijelaskan sebagai berikut:

1. **Input teks.** Teks seperti komentar, posting, atau transkrip video diubah menjadi bentuk numerik melalui embedding kata (*word embeddings*), sehingga kata-kata memiliki

representasi vektor berdimensi tinggi yang mempertahankan makna semantik.

2. **Word Embedding.** Teknik seperti Word2Vec, GloVe, atau embedding dari model Transformer digunakan untuk mengubah kata menjadi vektor angka. Contohnya, kata "judi" dan "slot" akan memiliki vektor yang lebih dekat secara matematis dibanding kata yang tidak berhubungan.
3. **RNN Layer.** RNN memproses input embedding secara berurutan, mempertahankan memori dari langkah sebelumnya. Setiap output pada langkah waktu *t* bergantung pada input saat itu dan memori dari langkah waktu sebelumnya (*hidden state*). Struktur ini membuat RNN mampu mengenali konteks dan pola panjang.
   Contoh: dalam kalimat "Jangan pernah main judi, karena merugikan", RNN dapat mengenali bahwa kata "judi" adalah bagian dari konteks negatif berdasarkan kata-kata sebelumnya.
4. **Fully Connected Layer.** Semua fitur yang dihasilkan oleh RNN digabungkan dan diproses melalui lapisan fully connected untuk membuat prediksi akhir.
5. **Output Layer.** Lapisan ini menghasilkan klasifikasi akhir, misalnya konten berbahaya vs normal, ujaran kebencian vs netral, atau konten promosi judi vs non-promosi.

Mekanisme Kerja RNN secara Teknis

RNN memiliki struktur loop yang memungkinkan *hidden state* dari satu langkah waktu digunakan sebagai input pada langkah waktu berikutnya. Secara matematis, proses ini dapat dijelaskan sebagai:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = g(W_{hy}h_t + b_y)$$

Di mana:

- x_t = input pada langkah waktu *t*
- h_t = hidden state pada langkah waktu *t*
- W_{xh}, W_{hh}, W_{hy} = bobot jaringan

- b_h,b_y = bias
- *f*,*g* = fungsi aktivasi (misalnya tanh atau softmax)

Mekanisme ini membuat RNN mampu mempertahankan informasi kontekstual dalam memproses teks panjang, yang menjadi kelemahan utama CNN.

Arsitektur RNN (sederhana)

1. **Input Layer.** Data berupa teks yang diubah menjadi representasi numerik (*word embeddings*).
2. **RNN Layer.** Lapisan ini memproses data sekuensial dengan mempertimbangkan langkah waktu sebelumnya. Variasi RNN populer termasuk:
    - Vanilla RNN
    - LSTM (Long Short-Term Memory): menangani masalah *vanishing gradient* untuk konteks panjang.
    - GRU (Gated Recurrent Unit): varian yang lebih sederhana dari LSTM.
3. **Fully Connected Layer.** Menggabungkan fitur dari RNN untuk menghasilkan representasi akhir.
4. **Output Layer.** Menghasilkan prediksi klasifikasi.

Contoh Kasus RNN dalam Moderasi Konten

1. **Moderasi Teks:** RNN dapat digunakan untuk mendeteksi ujaran kebencian dalam komentar media sosial. Misalnya, komentar **"Saya tidak suka dia karena ulahnya"** perlu dipahami secara kontekstual. RNN akan membaca kata demi kata, mempertimbangkan urutan dan makna kata sebelumnya, sehingga prediksi menjadi lebih akurat.
2. **Moderasi Audio:** RNN dapat memproses transkrip percakapan untuk mendeteksi konten negatif atau ujaran kasar. Misalnya, analisis urutan kata dalam percakapan telepon call center.
3. **Moderasi Video:** RNN dapat digunakan untuk memproses urutan frame atau transkrip subtitle video, sehingga mampu

mendeteksi adegan yang mengandung konten terlarang atau ujaran kebencian.

Tabel 4. 6 Kelebihan dan Keterbatasan RNN dalam Moderasi Konten

Aspek	Penjelasan
Kelebihan	- Mampu memahami konteks panjang dalam teks - Cocok untuk data sekuensial seperti teks, audio, atau video - Dapat mengenali pola tergantung urutan kata - Efektif untuk bahasa alami yang kompleks
Keterbatasan	- Pelatihan lebih lambat dibanding CNN - Sulit memproses data dalam jumlah besar secara paralel - Rentan terhadap masalah *vanishing gradient* (meskipun mitigasi dengan LSTM/GRU) - Membutuhkan dataset sekuensial yang baik

4.2.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) merupakan salah satu arsitektur Recurrent Neural Network (RNN) yang dikembangkan oleh Hochreiter dan Schmidhuber pada tahun 1997. LSTM hadir untuk mengatasi kelemahan utama RNN klasik, yaitu kesulitan mengingat informasi jangka panjang akibat fenomena *vanishing gradient*.

Dalam RNN biasa, informasi yang datang lebih awal dalam sebuah kalimat cenderung "hilang" atau tereduksi saat sampai ke bagian akhir. Hal ini menyulitkan model untuk memahami konteks yang panjang. LSTM memperbaiki hal tersebut dengan memperkenalkan mekanisme memori khusus yang memungkinkan jaringan untuk menyimpan, melupakan, dan menggunakan informasi secara selektif.

Struktur Inti LSTM

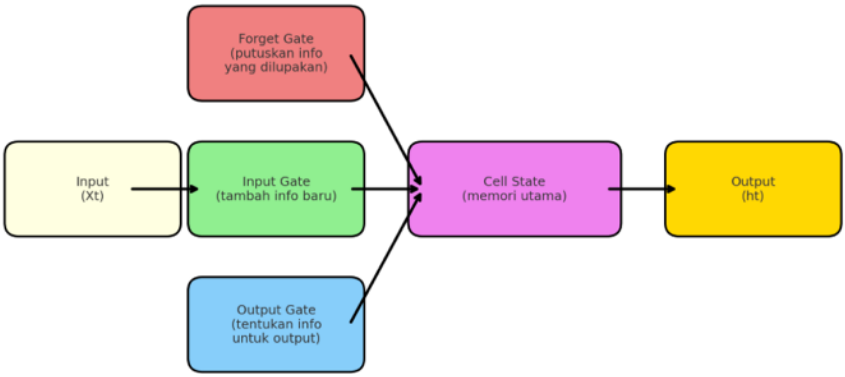
Unit dasar LSTM terdiri atas cell state (jalur memori utama) dan tiga komponen gerbang (*gates*) yang berfungsi sebagai pengatur aliran informasi. Setiap gerbang memiliki peran khusus dalam

memutuskan apakah suatu informasi perlu disimpan, dilupakan, atau digunakan untuk menghasilkan output.

1. **Cell State (Memori Utama)**
   - Merupakan jalur **memori jangka panjang** yang membawa informasi sepanjang urutan teks.
   - Disebut seperti "rel kereta" karena informasi dapat mengalir dari satu langkah ke langkah berikutnya dengan relatif sedikit perubahan.
   - Menjadi kunci utama LSTM dalam menjaga konteks yang panjang.
2. **Forget Gate (Gerbang Penghapus)**
   - Bertugas memutuskan informasi apa dari *cell state* sebelumnya yang harus dilupakan karena tidak relevan lagi.
   - Contoh: pada kalimat *"Saya tidak suka makan bakso"*, forget gate mungkin mengabaikan kata "makan" jika fokus klasifikasi hanya pada perasaan "tidak suka".
3. **Input Gate (Gerbang Pemasukan)**
   - Menentukan informasi baru apa yang perlu dimasukkan ke memori.
   - Informasi penting, seperti kata emosional (*"benci"*) atau kata kunci (*"hoaks"*), akan ditambahkan ke cell state sehingga tetap teringat dalam proses berikutnya.
4. **Output Gate (Gerbang Keluaran)**
   - Mengontrol informasi apa yang diambil dari memori untuk menghasilkan representasi keluaran pada langkah saat ini.
   - Representasi inilah yang digunakan untuk membuat prediksi, misalnya mengklasifikasikan apakah sebuah kalimat merupakan ujaran kebencian atau konten netral.

Dengan struktur ini, LSTM memiliki keunggulan utama: ia dapat menyimpan informasi penting untuk jangka panjang sekaligus membuang informasi yang tidak relevan, membuatnya jauh lebih

handal dibanding RNN biasa dalam memahami teks yang panjang dan kompleks.



Gambar 4. 3 Mekanisme Long Short-Term Memory (LSTM)

Diagram ini menggambarkan bagaimana LSTM mengelola aliran informasi dalam sebuah urutan teks.

1. **Input (X_t)** adalah kata atau token yang sedang diproses pada langkah tertentu (misalnya kata ke-3 dalam kalimat). Input ini diubah menjadi vektor numerik melalui embedding sebelum masuk ke LSTM.
2. **Forget Gate** memutuskan informasi dari *cell state* sebelumnya yang harus dibuang karena tidak relevan. Misalnya, kata "kemarin" bisa dilupakan jika tidak lagi penting untuk memahami arti kalimat saat ini.
3. **Input Gate** menentukan informasi baru dari kata yang sedang diproses untuk ditambahkan ke memori. Contoh: pada kalimat *"berita ini adalah hoaks"*, kata "hoaks" akan disimpan dalam memori karena krusial untuk klasifikasi.
4. **Cell State (Memori Utama)** membawa informasi sepanjang urutan kata. Dengan bantuan forget dan input gate, cell state hanya menyimpan informasi yang relevan. Hal ini membuat LSTM lebih unggul daripada RNN klasik, yang sering kehilangan informasi penting dari awal kalimat.
5. **Output Gate** mengontrol informasi apa yang diambil dari cell state untuk menghasilkan prediksi pada langkah tersebut.
6. **Output (h_t)** adalah representasi konteks terkini yang digunakan untuk tugas klasifikasi, misalnya apakah teks ini termasuk ujaran kebencian, hoaks, atau konten normal.

Dengan mekanisme ini, LSTM dapat mempertahankan konteks yang panjang sekaligus menyingkirkan informasi tidak relevan, sehingga lebih efektif dalam moderasi teks kompleks seperti artikel panjang, percakapan, atau kalimat yang mengandung sarkasme.

Proses Kerja LSTM dalam Moderasi Konten

1. **Input Teks**. Kalimat atau komentar dimasukkan ke dalam model dalam bentuk vektor (*word embeddings*).
2. **Pengolahan Berurutan**
    - Kata pertama diproses, lalu hidden state dan cell state diperbarui.
    - Setiap kata berikutnya diproses sambil membawa memori dari kata sebelumnya.
3. **Manajemen Memori dengan Gates**
    - Forget gate memutuskan apakah informasi lama masih relevan.
    - Input gate menambahkan informasi baru ke dalam memori.
    - Output gate menghasilkan representasi konteks yang diperbarui untuk prediksi.
4. **Prediksi Akhir**. Setelah seluruh kata diproses, output terakhir digunakan untuk mengklasifikasikan teks, misalnya "konten normal" atau "konten berbahaya".

Contoh Penerapan LSTM

1. **Deteksi Sarkasme**
    - Kalimat: *"Wah, pinter sekali ya, bikin hoaks tiap hari"*.
    - Kata "pinter sekali" seolah positif, tetapi makna keseluruhan menjadi negatif setelah ditambah "bikin hoaks tiap hari".
    - LSTM bisa menangkap pergeseran makna ini karena mampu mengingat konteks jangka panjang.
2. **Moderasi Artikel Hoaks**. Artikel berita palsu sering panjang dan rumit. LSTM dapat mempertahankan konteks dari awal hingga akhir teks, sehingga lebih andal dibanding RNN biasa.
3. **Moderasi Audio atau Percakapan.** Dalam analisis percakapan, LSTM bisa melacak topik yang sedang dibicarakan meskipun percakapan berlangsung lama.

Contoh Proses LSTM pada Kalimat: "*Berita ini adalah hoaks*"

1. Input Kata Pertama: "*Berita*"
    - Forget Gate: Tidak ada informasi sebelumnya yang perlu dilupakan (karena ini kata pertama).
    - Input Gate: Kata "Berita" dianggap relevan, sehingga dimasukkan ke memori (*cell state*).
    - Output Gate: Menghasilkan representasi awal yang menyatakan topik kalimat adalah "berita".
2. Input Kata Kedua: "*ini*"
    - Forget Gate: Memutuskan tetap menyimpan informasi tentang kata "berita".
    - Input Gate: Kata "ini" ditambahkan karena berfungsi memperjelas subjek kalimat.
    - Output Gate: Keluaran kini menunjukkan topik "berita ini".
3. Input Kata Ketiga: "*adalah*"
    - Forget Gate: Menyimpan informasi penting sebelumnya ("berita ini").
    - Input Gate: Menambahkan kata "adalah" yang menandakan adanya definisi atau klaim.
    - Output Gate: Representasi sekarang menyimpan konteks bahwa kalimat akan menjelaskan sesuatu tentang "berita ini".
4. Input Kata Keempat: "*hoaks*"
    - Forget Gate: Tidak menghapus informasi utama (subjek masih "berita ini").
    - Input Gate: Kata "hoaks" sangat penting → dimasukkan penuh ke dalam memori.
    - Output Gate: Sistem menekankan bahwa inti kalimat adalah klaim "berita ini adalah hoaks".

Hasil Akhir

Setelah seluruh kata diproses, cell state LSTM telah menyimpan konteks lengkap dari kalimat. Output terakhir dapat digunakan oleh model untuk melakukan klasifikasi konten → misalnya, sistem mendeteksi bahwa teks ini termasuk misinformasi/hoaks.

Tabel 4. 7 Ilustrasi Singkat Mekanisme Gates dalam Contoh Kalimat

Kata Masukan	**Forget Gate**	**Input Gate**	**Output Gate**	**Hasil Konteks Utama**
"Berita"	Tidak ada yang dihapus	Disimpan	Fokus awal	Topik: berita
"ini"	Menyimpan kata sebelumnya	Disimpan	Diperluas	"berita ini"
"adalah"	Menyimpan "berita ini"	Ditambahkan	Konteks klaim	"berita ini adalah"
"hoaks"	Menyimpan semua konteks	Disimpan penuh	Ditekankan	Klaim: "berita ini adalah hoaks"

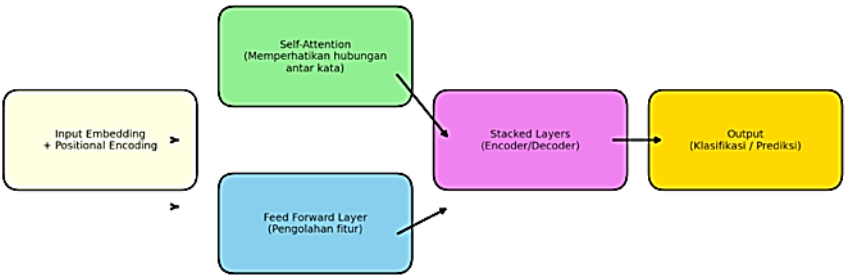
Tabel 4. 8 Kelebihan dan Keterbatasan LSTM dalam Moderasi Konten

Aspek	**Penjelasan**
Kelebihan	- Mampu mengingat konteks jangka panjang - Lebih akurat dibanding RNN klasik pada teks kompleks - Cocok untuk analisis teks panjang, artikel, atau percakapan
Keterbatasan	- Komputasi lebih berat dibanding RNN - Sulit dioptimalkan, butuh banyak data - Masih kalah efisien dibanding Transformer

4.2.4 Transformer

Transformer adalah arsitektur jaringan saraf modern yang diperkenalkan oleh Vaswani et al., (2017) dalam makalah berjudul *"Attention is All You Need"*. Berbeda dengan CNN atau RNN yang memproses data secara lokal atau berurutan, Transformer menggunakan mekanisme self-attention untuk memahami hubungan antar kata dalam sebuah kalimat secara paralel.

Dengan self-attention, model dapat memperhatikan kata-kata penting di seluruh kalimat sekaligus, tidak peduli apakah kata itu muncul di awal atau akhir. Hal ini membuat Transformer sangat kuat dalam memahami konteks panjang.



Gambar 4. 4 Cara Kerja Transformer

Cara Kerja Transformer (secara sederhana)

1. **Input Teks** → setiap kata diubah menjadi vektor (*embedding*).
2. **Positional Encoding** → ditambahkan agar model mengetahui urutan kata, karena Transformer tidak memproses kata secara sekuensial seperti RNN.
3. **Self-Attention Mechanism** → model menghitung seberapa penting setiap kata terhadap kata lain dalam kalimat. Misalnya, pada kalimat *"Dia tidak suka makan bakso"*, kata "tidak" akan sangat berpengaruh pada interpretasi kata "suka".
4. **Feed Forward Layer** → hasil perhatian diproses lebih lanjut untuk memperkuat representasi.
5. **Stacked Layers** → proses ini diulang berkali-kali sehingga model membangun pemahaman mendalam.
6. **Output Layer** → menghasilkan prediksi, misalnya klasifikasi apakah teks termasuk ujaran kebencian atau hoaks.

Contoh Kasus Transformer dalam Moderasi Konten

- Deteksi Ujaran Kebencian Multibahasa: Transformer dapat menangani teks dengan *code-switching* seperti *"Main slot jp bgt guys, langsung wd ke ovo"*, di mana campuran bahasa Indonesia-Inggris tetap bisa dipahami.
- Deteksi Hoaks Panjang: Artikel panjang yang membingungkan dapat dipahami lebih baik karena Transformer dapat melihat hubungan antar kata di seluruh teks.
- Moderasi Konten Multimodal: Dengan pengembangan model seperti CLIP atau Vision Transformer (ViT), Transformer dapat digunakan untuk memahami gambar, teks, bahkan kombinasi keduanya.

Tabel 4. 9 Kelebihan dan Keterbatasan Transformer dalam Moderasi Konten

Aspek	Penjelasan
Kelebihan	- Memahami konteks global dalam teks panjang - Efisien untuk pelatihan paralel - Dasar model bahasa besar (BERT, GPT, IndoBERT) - Dapat digunakan untuk teks, gambar, bahkan multimodal
Keterbatasan	- Membutuhkan dataset sangat besar - Komputasi mahal (butuh GPU/TPU besar) - Sulit dijelaskan (black box) - Potensial bias dari data latih yang tidak seimbang

4.3 AutoML

AutoML (*Automated Machine Learning*) adalah pendekatan yang dirancang untuk mengotomatisasi seluruh siklus pembangunan model machine learning maupun deep learning. Proses konvensional biasanya membutuhkan keahlian khusus: seorang data scientist harus melakukan feature engineering, memilih algoritma yang sesuai, menyesuaikan hyperparameter, serta menguji performa model satu per satu. Tahapan ini memakan banyak waktu, tenaga, dan keahlian.

AutoML menyederhanakan proses tersebut dengan membuat sistem yang secara otomatis dapat:

1. Menganalisis data input: apakah berupa teks, gambar, audio, atau gabungan (multimodal).
2. Melakukan feature engineering otomatis: misalnya, mengubah teks menjadi *word embeddings* (Word2Vec, FastText, BERT embeddings) tanpa campur tangan manual.
3. Memilih algoritma terbaik: mencoba berbagai kandidat, mulai dari *klasik* (Naive Bayes, SVM, Random Forest) hingga *modern* (CNN, RNN, Transformer).
4. Menyetel hyperparameter secara otomatis: misalnya, menentukan *learning rate*, jumlah layer, atau jumlah pohon dalam Random Forest.

5. Mengevaluasi performa dengan metrik yang relevan: akurasi, precision, recall, F1-score.
6. Membangun ensemble model (gabungan beberapa model terbaik) untuk hasil yang lebih andal.

Contoh Kasus AutoML dalam Moderasi Konten Digital

Kasus: Deteksi Iklan Judi Online di Media Sosial Indonesia

- **Data**: ribuan unggahan teks (caption Instagram, postingan Facebook), gambar (banner promosi), dan video pendek (iklan terselubung).
- **Tantangan**:
  - ⇒ Bahasa yang digunakan sering *disamarkan*, misalnya menulis "jvdi" alih-alih "judi".
  - ⇒ Promosi berbentuk gambar dengan teks di dalamnya (perlu OCR untuk ekstraksi).
  - ⇒ Konten video/audio dengan kata kunci kode (*slang*) yang berubah-ubah.

Tanpa AutoML: Tim riset harus mencoba satu per satu algoritma (misalnya CNN untuk gambar, LSTM untuk teks), menyetel hyperparameter, dan menguji performa manual. Proses ini bisa memakan waktu berminggu-minggu.

Dengan AutoML:

- Peneliti cukup memberikan dataset berlabel: *normal* vs *judi online*.
- AutoML secara otomatis:
  - ⇒ Menguji berbagai model (SVM, Random Forest, CNN, Transformer).
  - ⇒ Mengoptimalkan parameter (misalnya learning rate pada CNN, kernel pada SVM).
  - ⇒ Membandingkan hasil tiap model.
  - ⇒ Menggabungkan model terbaik dalam ensemble: misalnya, CNN untuk deteksi visual + Transformer untuk

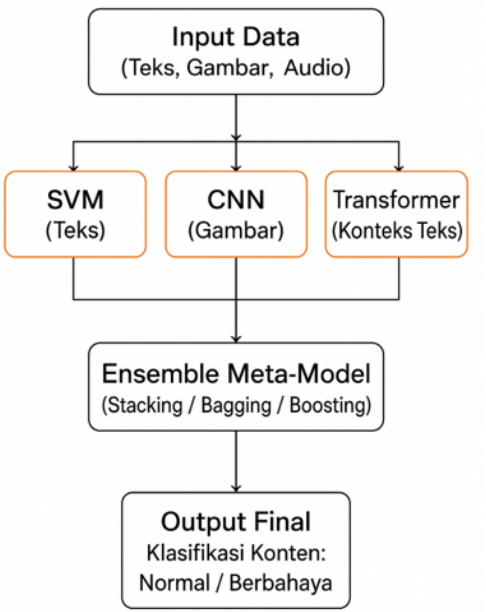
analisis teks → digabung dengan *stacking* → hasil akhir lebih akurat.

Hasilnya: AutoML mungkin menghasilkan model ensemble yang memiliki F1-score lebih tinggi dibanding model tunggal. Misalnya:

- CNN sendiri: 85% F1-score.
- Transformer sendiri: 88% F1-score.
- Ensemble CNN + Transformer: **91% F1-score**.

Hal ini menunjukkan bahwa ensemble otomatis dalam AutoML mampu mengurangi kelemahan masing-masing model dan memberikan kinerja lebih konsisten.

Jadi, AutoML bukan sekadar "shortcut" untuk membangun AI, tetapi sebuah sistem yang mempercepat eksperimen, memilih model terbaik secara obyektif, dan menghasilkan kombinasi model (ensemble) yang lebih tahan terhadap variasi data di dunia nyata. Untuk memahami alur kerja dari sistem AutoML ini, dapat dilihat pada Gambar XX yang menampilkan proses klasifikasi konten dari awal hingga akhir. Tahap pertama dimulai dengan input data yang bisa berupa teks, gambar, atau audio, kemudian masing-masing jenis data diproses oleh model yang sesuai: SVM untuk teks sederhana, CNN untuk gambar, dan Transformer untuk teks dengan konteks yang lebih kompleks. Hasil dari setiap model tersebut selanjutnya digabungkan melalui sebuah ensemble meta-model dengan pendekatan seperti stacking, bagging, atau boosting untuk meningkatkan akurasi dan keandalan. Pada akhirnya, sistem ini menghasilkan output final berupa klasifikasi konten, apakah termasuk kategori normal atau berbahaya.



Gambar 4. 5 Proses Klasifikasi Konten dengan AutoML

Berikut penjelasan dari gambar di atas yang menggambarkan alur kerja sistem klasifikasi konten berbasis AutoML.

1. Input Data (Teks, Gambar, Audio)
   - Sistem menerima berbagai macam data mentah sebagai masukan, meliputi:
     - Teks: artikel, komentar, chat, caption, dokumen.
     - Gambar: foto, ilustrasi, tangkapan layar.
     - Audio: percakapan suara, rekaman.
   - Data ini adalah bahan utama yang perlu diproses agar bisa diklasifikasikan.

2. Pemrosesan dengan Model Spesifik

Input data masuk ke jalur model yang sesuai dengan karakteristiknya:

- SVM (Support Vector Machine) → khusus Teks
  - SVM digunakan untuk melakukan klasifikasi berbasis fitur teks.
  - Biasanya setelah teks diolah (tokenisasi, ekstraksi fitur seperti TF-IDF atau word embedding).
  - Cocok untuk teks sederhana dengan dataset terbatas.
- CNN (Convolutional Neural Network) → khusus Gambar

    - CNN digunakan untuk menganalisis data visual.
    - Menangkap pola seperti tepi, bentuk, dan objek pada gambar.
    - Cocok untuk mendeteksi konten visual yang berbahaya atau normal.
- Transformer → khusus Konteks Teks
    - Digunakan untuk memahami teks dengan konteks yang lebih kompleks.
    - Contohnya model BERT, GPT, atau sejenisnya.
    - Mampu memahami hubungan antar kata dalam kalimat/paragraf sehingga dapat menangkap makna lebih dalam dibandingkan SVM.

3. Ensemble Meta-Model (Stacking / Bagging / Boosting)
    - Hasil prediksi dari masing-masing model (SVM, CNN, Transformer) tidak langsung dipakai, tapi digabung dengan pendekatan ensemble.
    - Stacking: menggabungkan prediksi beberapa model sebagai input bagi model meta.
    - Bagging: melakukan voting rata-rata (misalnya Random Forest).
    - Boosting: memperkuat model lemah menjadi lebih kuat dengan bobot tertentu (misalnya XGBoost, AdaBoost).
    - Tujuannya: meningkatkan akurasi dan robustnes sistem, sehingga kesalahan dari satu model bisa dikompensasi oleh model lain.
4. Output Final
    - Setelah melalui ensemble, sistem menghasilkan keputusan akhir:
        - Normal → konten tidak mengandung bahaya, bisa dipublikasikan atau disimpan.
        - Berbahaya → konten mengandung potensi risiko, misalnya ujaran kebencian, pornografi, ancaman, atau hoaks.
    - Output ini adalah hasil klasifikasi yang lebih kuat karena memadukan keunggulan berbagai model.

Inti dari diagram ini yaitu sistem dirancang untuk menerima berbagai jenis input (multimodal), memprosesnya dengan model yang sesuai (SVM untuk teks sederhana, CNN untuk gambar, Transformer untuk teks kontekstual), lalu menggabungkan hasilnya dengan ensemble untuk menghasilkan keputusan klasifikasi yang lebih akurat dan andal.

4.4 Transfer Learning dan Model Bahasa (BERT, IndoBERT)

4.4.1 Konsep Dasar Transfer Learning

Dalam machine learning tradisional, model harus dilatih dari nol dengan dataset yang besar agar dapat bekerja optimal. Namun, di dunia nyata — terutama dalam moderasi konten digital — sering kali dataset terbatas dan variatif. Di sinilah transfer learning berperan. Intinya adalah memanfaatkan pengetahuan model besar yang sudah dilatih sebelumnya pada dataset raksasa, lalu menyesuaikannya (fine-tuning) untuk tugas spesifik.

Analogi sederhana: Bayangkan seseorang yang sudah mahir bahasa Inggris. Ketika belajar bahasa Indonesia, ia tidak mulai dari nol, melainkan hanya menyesuaikan kosakata dan tata bahasa baru. Begitu juga dengan model pre-trained: ia sudah menguasai "struktur bahasa" umum, sehingga lebih cepat beradaptasi dengan domain baru.

4.4.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT adalah model berbasis Transformer yang dikembangkan Google (2018). Keunggulannya ada pada kemampuan memahami konteks dua arah (Devlin et al., 2019).

Cara kerja BERT:

- Input kalimat diubah menjadi token (potongan kata).
- Token diberi embedding dan positional encoding.
- Self-Attention digunakan untuk melihat hubungan antar token di seluruh kalimat.

- Output berupa representasi yang mempertimbangkan konteks global.

📌 Contoh 1: Ujaran Kebencian Terselubung

- Kalimat: *"Orang itu pintar banget, sampai-sampai bisa bikin hoaks tiap hari."*
- Model klasik (SVM, Naive Bayes): mungkin menilai kalimat ini netral karena tidak ada kata kasar eksplisit.
- BERT: memahami bahwa "pintar" di sini digunakan secara sarkastik, sehingga bisa mendeteksi konten negatif.

📌 Contoh 2: Ambiguitas Kata

- Kalimat: *"Bank di tepi sungai itu indah sekali."*
- Kata "bank" di sini bukan lembaga keuangan.
- BERT mampu memahami makna dengan melihat keseluruhan kalimat, bukan hanya kata tunggal.

4.4.3 IndoBERT (Varian Bahasa Indonesia)

IndoBERT adalah adaptasi BERT yang dilatih menggunakan korpus teks bahasa Indonesia (Wikipedia, berita, media sosial, dsb.) (Koto et al., 2020). Model ini lebih sesuai dengan kebutuhan lokal karena:

- Bahasa Indonesia sering menggunakan slang, singkatan, dan campuran bahasa (code-switching).
- Struktur kalimat Indonesia berbeda dari bahasa Inggris.

📌 Contoh 3: Bahasa Gaul & Campuran

- Kalimat: *"Lo bener2 noob parah, jangan sok Inggris deh bro."*
- Model bahasa Inggris → gagal, karena kata "noob" dan "sok Inggris" tidak umum.
- IndoBERT → lebih paham konteks bahwa kalimat ini merendahkan orang lain.

📌 Contoh 4: Deteksi Judi Online

- Kalimat: *"Join room 88, dapet chip gratis, aman 100% bro."*
- Kata "room 88" dan "chip gratis" sering menjadi kode promosi judi.
- IndoBERT dapat belajar pola ini dari data lokal dan mengenalinya sebagai promosi ilegal.

📌 Contoh 5: Hoaks dengan Bahasa Gaul

- Kalimat: *"Fix banget vaksin bikin chip, bro konspirasi nih."*
- Model klasik mungkin gagal karena kata "chip" bisa bermakna teknologi.
- IndoBERT lebih akurat karena memahami konteks *"vaksin + chip + konspirasi"* sebagai hoaks.

Tabel 4. 10 Perbandingan BERT dan IndoBERT untuk Moderasi Konten

Aspek	**BERT (Bahasa Inggris)**	**IndoBERT (Bahasa Indonesia)**
Dataset Pra-latih	Wikipedia + BookCorpus (bahasa Inggris)	Korpus berita, media sosial, Wikipedia Indonesia
Kekuatan	Pemahaman konteks global, banyak varian (RoBERTa, DistilBERT)	Lebih akurat untuk teks bahasa Indonesia dan slang
Keterbatasan	Tidak peka pada slang/gaul lokal	Dataset pra-latih lebih kecil dibanding BERT global
Aplikasi Moderasi	Analisis teks global, riset multibahasa	Deteksi ujaran kebencian, hoaks, judi online di Indonesia

Dengan uraian ini, maka akan mudah memahami bahwa transfer learning dengan model BERT dan varian lokal seperti IndoBERT bukan sekadar teknologi tambahan, melainkan fondasi utama (tulang punggung) dalam sistem moderasi konten digital modern. Hal ini karena:

1. Bahasa media sosial sangat dinamis – penuh slang, singkatan, dan campuran bahasa (*code-switching*) yang sulit dipahami

algoritma klasik. Model berbasis Transformer seperti BERT mampu menangkap konteks kompleks ini.

2. Data lokal sangat penting – di Indonesia, istilah terkait hoaks, ujaran kebencian, atau promosi judi memiliki ciri khas tersendiri. IndoBERT yang dilatih dengan korpus bahasa Indonesia memiliki keunggulan dalam menangkap pola lokal tersebut.
3. Transfer learning mempercepat adaptasi – alih-alih membangun model dari nol (yang mahal dan lambat), BERT/IndoBERT bisa langsung *fine-tuned* dengan dataset moderasi yang relatif kecil, sehingga sistem lebih cepat siap digunakan.
4. Performa terbukti unggul – berbagai penelitian menunjukkan bahwa model BERT dan IndoBERT secara konsisten mengungguli algoritma klasik (Naive Bayes, SVM, Random Forest) maupun deep learning konvensional (RNN, LSTM) dalam tugas klasifikasi teks berbahasa Indonesia.
5. Skalabilitas moderasi konten – karena mampu dipakai lintas domain (hoaks, ujaran kebencian, pornografi, judi online, dsb.), BERT/IndoBERT bisa menjadi basis tunggal yang serbaguna dalam sistem moderasi konten berskala besar.

Dengan kata lain, tanpa transfer learning, moderasi konten di media sosial akan sulit mengimbangi laju produksi konten berbahaya yang terus berkembang. Oleh karena itu, model seperti BERT dan IndoBERT adalah tulang punggung yang memungkinkan moderasi konten berjalan cepat, adaptif, dan akurat di ekosistem digital Indonesia.

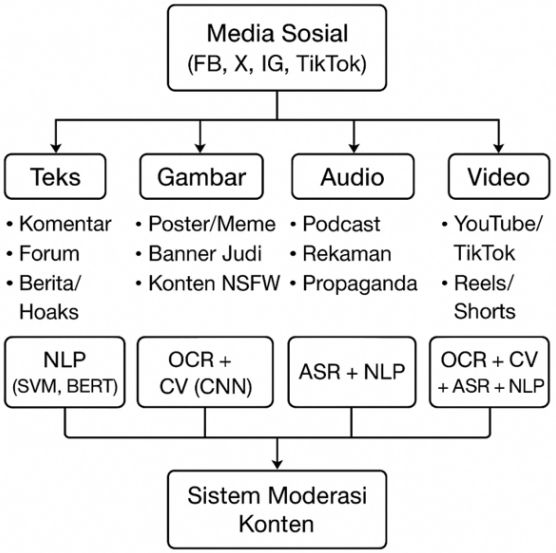
BAB V
DATA UNTUK MODERASI KONTEN

Salah satu prinsip utama dalam pengembangan sistem kecerdasan buatan adalah ungkapan: “garbage in, garbage out.” Artinya, kualitas hasil model sangat bergantung pada kualitas data yang digunakan dalam pelatihannya. Dalam konteks moderasi konten digital, tantangan terbesar justru terletak pada pengumpulan, anotasi, dan pengelolaan data. Data moderasi tidak hanya berupa teks sederhana, tetapi juga mencakup gambar, audio, dan video yang sering mengandung konten berbahaya atau terselubung. Oleh karena itu, Bab ini akan membahas:

1. Sumber data moderasi konten digital (teks, gambar, audio, video).
2. Proses anotasi dan labeling, termasuk peran manusia dalam memberi label yang konsisten.
3. Tantangan linguistik dan teknis, seperti slang, code-switching, serta kesalahan OCR/ASR.
4. Etika penggunaan data, mengingat data moderasi sering kali mengandung konten sensitif.

5.1 Sumber Data Moderasi Konten

Kualitas sistem moderasi konten berbasis kecerdasan buatan sangat ditentukan oleh sumber data yang digunakan dalam pelatihannya. Data ini bisa berasal dari berbagai kanal digital dan memiliki bentuk yang berbeda: teks, gambar, audio, maupun video. Masing-masing jenis data memiliki tantangan tersendiri dalam pengolahan dan anotasinya.



Gambar 5. 1 Diagram Alur Sistem Moderasi Konten Media Sosial

Dengan memahami berbagai sumber data, peneliti dan praktisi dapat membangun sistem moderasi konten yang lebih komprehensif. Sistem tidak hanya fokus pada teks, tetapi juga mampu mendeteksi konten berbahaya yang tersembunyi dalam gambar, audio, maupun video. Berikut penjelasannya dari Gambar 5.1.

1. Data Teks

Teks adalah bentuk data yang paling banyak digunakan dalam moderasi konten. Contohnya berasal dari:

- Komentar media sosial seperti Facebook, Instagram, dan Twitter/X. Komentar ini bisa berisi ujaran kebencian, spam, atau iklan judi.
- Forum online seperti Reddit, Kaskus, atau forum ilegal tempat diskusi perjudian berlangsung.
- Artikel berita dan blog yang digunakan sebagai pembanding antara konten faktual vs hoaks.

Sebagai contoh, komentar *"Join room 88 dapet chip gratis bro!"* dapat dilabeli sebagai iklan judi online karena mengandung kata kunci "*room 88*" dan "*chip gratis*" yang sering dipakai untuk promosi perjudian.

2. Data Gambar

Banyak konten berbahaya di media sosial berbentuk visual, bukan teks.

- Poster atau banner sering digunakan untuk promosi judi, misalnya gambar slot machine dengan tulisan "*deposit 10k dapat bonus 50k*".
- Meme yang berisi teks + gambar bisa menjadi sarana penyebaran ujaran kebencian dengan cara satir atau sindiran visual.

Data gambar biasanya dikumpulkan melalui web scraping dari media sosial atau menggunakan dataset publik seperti *NSFW image dataset*. Untuk menganalisisnya, sistem menggunakan Computer Vision serta OCR (Optical Character Recognition) untuk membaca teks dalam gambar. Sebagai contoh: gambar meme dengan teks *"Bangsa ini hancur gara-gara si X"* → bisa dikategorikan sebagai ujaran kebencian.

3. Data Audio

Konten audio semakin penting karena berkembangnya platform berbasis suara dan podcast.

- Rekaman percakapan dapat berisi ujaran kebencian, ajakan radikalisme, atau propaganda.
- Podcast dan audio dari video kadang digunakan untuk menyebarkan misinformasi secara terselubung.

Data audio biasanya diproses menggunakan Automatic Speech Recognition (ASR) untuk mengubah ucapan menjadi teks, sehingga bisa dianalisis dengan algoritma NLP. Contoh: rekaman suara *"Bro, jangan percaya pemerintah, vaksin itu racun!"* → setelah ditranskrip, bisa dikategorikan sebagai hoaks.

4. Data Video

Video adalah bentuk konten yang paling kompleks karena menggabungkan teks, audio, dan gambar.

- Platform seperti YouTube, TikTok, dan Instagram Reels penuh dengan video pendek yang bisa mengandung ujaran kebencian, pornografi, atau promosi judi.

- Iklan terselubung sering muncul dalam bentuk hiburan singkat, misalnya video lucu dengan teks promo judi yang disisipkan.

Analisis video biasanya dilakukan dengan memecahnya menjadi beberapa komponen:

- Frame gambar untuk mendeteksi visual sensitif (pornografi, kekerasan).
- OCR (Optical Character Recognition) untuk membaca teks overlay atau subtitle.
- ASR (Automatic Speech Recognition) untuk mengubah audio menjadi teks yang bisa dianalisis lebih lanjut.

Contohnya, sebuah video TikTok menampilkan animasi dengan teks overlay *"join slot88 biar kaya raya"*. Dengan kombinasi OCR dan NLP, sistem dapat mengenali konten tersebut sebagai iklan judi.

5.2 Proses Anotasi dan Labeling

Agar sebuah model AI dapat belajar membedakan konten yang berbahaya dari yang normal, data harus diberi label terlebih dahulu. Proses pemberian label ini disebut anotasi. Dalam konteks moderasi konten digital, anotasi biasanya dilakukan oleh manusia (manual) atau semi-otomatis dengan bantuan sistem, lalu diperiksa ulang untuk memastikan kualitas.

1. Anotasi Manual oleh Manusia

Anotasi manual dilakukan dengan melibatkan tim anotator yang membaca, menonton, atau mendengarkan data, kemudian memberi label sesuai kategori yang telah ditentukan.

- Kategori umum: normal, hoaks, ujaran kebencian, pornografi, judi online, spam.
- Anotator perlu diberikan pedoman (annotation guideline) agar konsisten.

Contoh:

- Teks *"Join room 88 dapet chip gratis bro!"* → diberi label *judi online*.

- Meme dengan tulisan *"Kaum X memang nggak guna"* → diberi label *ujaran kebencian*.

Kelebihan anotasi manual adalah tingginya akurasi dalam memahami konteks, misalnya membedakan antara humor, sarkasme, atau ujaran serius. Namun, kekurangannya adalah mahal, memakan waktu lama, dan berisiko membuat anotator lelah atau trauma karena harus melihat konten berbahaya berulang kali.

2. Anotasi Semi-Otomatis

Karena volume data sangat besar, anotasi manual saja tidak cukup. Oleh karena itu, sering digunakan pendekatan semi-otomatis:

- Rule-based pre-labeling: sistem menggunakan kata kunci tertentu (misalnya "slot", "chip gratis", "deposit") untuk menandai data yang kemungkinan besar adalah iklan judi.
- Weak supervision: sistem menggunakan model awal (baseline) untuk memberi label kasar, lalu anotator manusia memverifikasi dan memperbaikinya.

Contoh: Jika sebuah sistem mendeteksi teks "ayo join slot 88, dapet bonus free chip" sebagai *judi online*, anotator hanya perlu mengonfirmasi label, bukan memberi label dari nol. Ini mempercepat proses labeling ribuan data.

3. Quality Control (Kontrol Kualitas Anotasi)

Untuk memastikan konsistensi label, biasanya dilakukan langkah berikut:

- Double annotation: setiap data diberi label oleh minimal dua anotator.
- Inter-annotator agreement (IAA): mengukur tingkat kesepakatan antar anotator.
    - Umumnya dihitung menggunakan Cohen's Kappa (dua anotator) atau Fleiss' Kappa (lebih dari dua anotator).
    - Nilai > 0.75 dianggap baik, nilai 0.4–0.6 dianggap sedang, nilai < 0.4 berarti perlu revisi pedoman anotasi.

Contoh: Teks *"Dasar lo bucin tolol, jangan ikut campur urusan orang!"*

- Anotator A → *ujaran kebencian*
- Anotator B → *bullying verbal*
- Jika tingkat kesepakatan rendah, dilakukan diskusi atau majority voting untuk memutuskan label final.

4. Tantangan Anotasi dalam Moderasi Konten

a) Ambiguitas Bahasa

Bahasa manusia penuh dengan nuansa yang bisa menimbulkan ambiguitas. Dalam anotasi moderasi konten, satu kalimat bisa memiliki interpretasi berbeda tergantung pada konteks sosial, situasi percakapan, atau nada bicara yang tidak selalu terlihat dalam teks. Sebagai contoh:

- *"Dasar anak itu nakal banget!"* → bisa bermakna candaan dalam lingkup keluarga atau bisa bermakna penghinaan jika ditujukan ke orang asing.
- *"Keren banget, tiap hari bikin hoax baru."* → bisa bermakna satire, bukan pujian.

Adapun tantangannya yaitu anotator harus berhati-hati membedakan antara candaan, satire, atau penghinaan serius. Tanpa pedoman yang jelas, hasil anotasi bisa berbeda-beda antar individu.

b) Slang & Singkatan

Media sosial dipenuhi bahasa gaul, singkatan, dan modifikasi ejaan. Hal ini sengaja dilakukan pengguna untuk mempercepat komunikasi atau menyamarkan maksud tertentu. Contoh:

- "jvdi" = judi, ditulis dengan huruf diganti angka/karakter untuk menghindari deteksi.
- "anjir" atau "anjay" = ekspresi emosi, bisa netral atau kasar tergantung konteks.
- "noob" = istilah dalam gaming yang berarti pemula, tetapi sering digunakan sebagai ejekan.

Tantangannya yaitu anotator harus memahami konteks budaya digital dan perkembangan istilah baru yang muncul hampir setiap bulan. Tanpa pemahaman ini, anotasi bisa salah kaprah.

c) Campuran Bahasa (*Code-Switching*)

Fenomena *code-switching* atau campur bahasa sangat umum di media sosial Indonesia. Pengguna sering mencampur bahasa Indonesia dengan Inggris, bahkan bahasa daerah, untuk mengekspresikan diri. Contoh:

- "Bro, fix ini hoax banget, jangan percaya deh."
- "Lu so toxic sih, mainnya nggak fair banget."
- "Cuma newbie doang kok, chill aja."

Tantangan: model klasik dan anotator baru bisa bingung apakah suatu istilah termasuk ujaran kebencian, bercanda, atau sekadar ekspresi gaul. Selain itu, *code-switching* menuntut anotator memiliki pemahaman lintas bahasa.

d) Beban Psikologis Anotator

Proses anotasi sering melibatkan konten berbahaya atau ekstrem, seperti kekerasan, pornografi, ujaran kebencian, atau propaganda. Paparan konten semacam ini berulang kali dapat menimbulkan tekanan mental, stres, hingga trauma psikologis bagi anotator. Contoh Konten:

- Gambar kekerasan fisik yang brutal.
- Video pornografi yang mengeksploitasi anak.
- Ujaran kebencian berbasis SARA yang intens.

Tantangan yaitu:

- Anotator bisa mengalami desensitisasi (jadi terbiasa melihat konten berbahaya).
- Bisa pula mengalami burnout atau trauma.
- Oleh karena itu, perusahaan teknologi besar biasanya menyediakan dukungan psikologis dan rotasi kerja untuk mengurangi dampak negatif ini.

Jadi, tantangan anotasi bukan hanya masalah teknis (slang, ambiguitas, code-switching), tetapi juga masalah kesehatan mental manusia yang terlibat dalam proses labeling. Oleh karena itu, anotasi harus dilakukan dengan pedoman jelas, pelatihan khusus, dan dukungan psikologis yang memadai.

5. Studi Kasus: Anotasi Ambigu dalam Moderasi Konten

Misalnya terdapat kalimat: *"Dasar bucin, hidup lo nggak ada arti!"*
Hasil Anotasi dari Tim, sebagai berikut:

Tabel 5. 1 Contoh Hasil Anotasi

Anotator	**Label**	**Alasan**
Anotator A	*Ujaran Kebencian*	Kata "hidup lo nggak ada arti" dianggap sebagai penghinaan serius terhadap martabat seseorang.
Anotator B	*Bullying Verbal (Ringan)*	Menilai bahwa ini lebih ke ejekan personal (bullying), bukan kebencian berbasis identitas kelompok (SARA).
Anotator C	*Candaan / Sarkasme*	Menganggap bahwa di konteks pertemanan atau percintaan, "bucin" (budak cinta) sering digunakan sebagai ejekan bercanda.

a) Analisis Perbedaan Interpretasi

1. Ambiguitas Konteks → Tanpa mengetahui hubungan antar pengguna (teman akrab vs orang asing), sulit menentukan apakah kalimat ini bercanda atau hinaan serius.
2. Slang “bucin” → Bisa netral (candaan ringan) atau ofensif (pelecehan emosional), tergantung konteks.
3. Kalimat tambahan “hidup lo nggak ada arti” → bagi sebagian anotator dianggap penghinaan serius, bagi yang lain dianggap bercanda hiperbolis.

b) Quality Control (Penyelesaian)

- Diskusi tim: Anotator membahas maksud kalimat, mempertimbangkan konteks sosial umum di media sosial.
- Pedoman Anotasi: Dalam guideline disebutkan bahwa jika sebuah teks menyerang martabat pribadi (meski bukan berbasis SARA), maka tetap dikategorikan sebagai *ujaran kebencian personal/bullying*.
- Keputusan Final: Label disepakati sebagai Bullying Verbal / Ujaran Kebencian Personal.

Cohen’s Kappa Score pada kasus ini rendah (misalnya sekitar 0.45), menunjukkan adanya perbedaan interpretasi signifikan antar anotator. Setelah revisi pedoman, konsistensi meningkat.

5.3 Tantangan dalam Pengumpulan dan Pemrosesan Data

Pengumpulan data untuk moderasi konten digital bukanlah tugas yang sederhana. Di satu sisi, data harus mencerminkan realitas di media sosial yang sangat dinamis. Di sisi lain, data yang diambil sering kali memiliki kualitas yang beragam dan mengandung noise. Berikut adalah beberapa tantangan utama yang sering dihadapi:

1. Slang dan Bahasa Gaul

Media sosial dipenuhi dengan kata-kata gaul, singkatan, atau variasi ejaan yang berbeda dari bahasa formal. Pengguna sengaja menggunakan istilah ini untuk mempercepat komunikasi atau bahkan untuk menghindari deteksi sistem otomatis.

- Contoh: kata "jvdi", "sl0t", atau "chip gratis" sering digunakan sebagai bentuk kamuflase iklan judi online.
- Tantangan: sistem NLP klasik sulit mengenali variasi kata ini, sehingga diperlukan model yang lebih adaptif atau teknik preprocessing khusus (misalnya normalisasi slang).

2. Code-Switching (Campuran Bahasa)

Fenomena *code-switching* sangat lazim di Indonesia, di mana pengguna mencampur bahasa Indonesia dengan bahasa Inggris (atau bahasa daerah).

- Contoh: *"Bro, fix ini hoax banget, don't trust it ya!"*
- Tantangan:
  - Model berbasis satu bahasa bisa gagal memahami maksud kalimat.
  - Membutuhkan model multilingual (seperti mBERT) atau model lokal (IndoBERT) yang sudah dilatih pada data campuran.

3. Kesalahan OCR (Optical Character Recognition) dan ASR (Automatic Speech Recognition)

Banyak konten berbahaya disamarkan dalam bentuk gambar berisi teks (misalnya poster judi) atau audio/video (misalnya

propaganda dalam rekaman suara). Untuk mengekstrak teks, sistem menggunakan OCR dan ASR, tetapi keduanya rentan salah.

- Contoh OCR: teks "join sl0t88 skrg" bisa dibaca salah menjadi "join slotss" tergantung font atau kualitas gambar.
- Contoh ASR: kalimat "chip gratis" dalam audio bisa ditranskripsi salah menjadi "cip gratis" atau "cheap gratis".
- Tantangan: kesalahan ekstraksi ini bisa menurunkan akurasi model moderasi karena input data menjadi tidak konsisten.

4. Ambiguitas Konteks

Tidak semua kalimat dengan kata kasar bermakna kebencian. Begitu pula tidak semua kalimat yang tampak netral bebas dari bahaya. Contohnya:

- *"Dia raja slot di kampungnya."* → bisa bermakna kiasan (pintar bermain), atau benar-benar iklan judi.
- *"Pintar banget, tiap hari bikin hoax baru."* → terdengar seperti pujian, tetapi sebenarnya sindiran negatif.

Tantangannya disini yaitu: anotator manusia pun bisa berbeda pendapat, sehingga perlu guideline jelas dan model yang bisa menangkap makna kontekstual.

5. Volume Data yang Sangat Besar

Setiap menit, jutaan konten baru dipublikasikan di berbagai platform. Mengumpulkan, menyaring, dan memberi label pada data dalam skala besar memerlukan infrastruktur kuat. Tantangannya:

- Tidak mungkin seluruh data dianotasi manual → butuh strategi sampling atau semi-otomatis.
- Harus menjaga representasi data agar tidak bias (misalnya tidak hanya fokus pada satu platform atau satu kelompok bahasa saja).

Dengan memahami tantangan ini, peneliti dan praktisi dapat merancang strategi pengumpulan serta preprocessing data yang lebih cermat. Misalnya, menggunakan kamus slang dinamis, mengembangkan model multilingual, serta memastikan adanya kontrol kualitas anotasi untuk mengurangi ambiguitas.

== Bagian III ==
APLIKASI MODERASI KONTEN DIGITAL

Bagian ini membahas implementasi nyata dari teknologi kecerdasan buatan dalam memoderasi berbagai jenis konten di media sosial. Bab keenam mengulas moderasi konten teks, seperti deteksi ujaran kebencian, identifikasi hoaks dan misinformasi, serta klasifikasi promosi ilegal. Bab ketujuh berfokus pada moderasi konten visual, meliputi pengenalan gambar eksplisit, penerapan OCR untuk mengekstraksi teks dari gambar atau video, serta analisis video dalam mendeteksi konten sensitif. Bab kedelapan melengkapi pembahasan dengan moderasi konten audio dan multimodal, termasuk penggunaan Automatic Speech Recognition (ASR), deteksi ujaran toksik, serta integrasi analisis lintas media antara teks, gambar, dan audio. Melalui bagian ini, pembaca dapat memahami bagaimana kecerdasan buatan digunakan secara praktis untuk menjaga ruang digital tetap aman, sehat, dan bermanfaat, dengan pendekatan yang komprehensif terhadap berbagai bentuk data.

BAB VI
MODERASI KONTEN TEKS

Setelah memahami dasar teori, algoritma populer, serta tantangan dalam data, langkah berikutnya adalah melihat bagaimana konsep tersebut diterapkan dalam kasus nyata. Moderasi konten digital dapat dibagi menjadi tiga domain utama: teks, visual (gambar & video), dan audio/multimodal. Teks merupakan bentuk konten digital yang paling banyak ditemukan di berbagai platform. Komentar pengguna, caption di media sosial, judul artikel, bahkan transkripsi dari video/audio semuanya berbasis teks. Karena sifatnya yang singkat, spontan, dan penuh variasi bahasa, teks di media sosial sering menjadi wadah bagi penyebaran ujaran kebencian, hoaks, maupun promosi ilegal. Moderasi teks menjadi fondasi utama dalam sistem deteksi otomatis. Sebelum sistem dapat memahami gambar, audio, atau video, teks biasanya menjadi titik masuk pertama karena lebih mudah diproses oleh algoritma berbasis Natural Language Processing (NLP).
Dalam bab ini, kita akan membahas tiga aplikasi utama:

1. Deteksi Ujaran Kebencian
2. Deteksi Hoaks dan Misinformasi
3. Deteksi Promosi Ilegal

6.1 Deteksi Ujaran Kebencian

Ujaran kebencian (*hate speech*) merupakan salah satu bentuk konten paling berbahaya di media sosial. Secara umum, ujaran kebencian dapat didefinisikan sebagai ekspresi. Misalnya ekspersi berupa teks, gambar, maupun audio dengan tujuan untuk menyerang atau merendahkan individu maupun kelompok tertentu berdasarkan atribut identitas, seperti ras, agama, gender, orientasi seksual, kewarganegaraan, maupun kondisi fisik. Bentuknya bisa berupa hinaan langsung, stereotipe negatif, ajakan diskriminasi, hingga provokasi yang mendorong kekerasan terhadap kelompok tertentu.

Deteksi ujaran kebencian membutuhkan pendekatan yang cerdas dan adaptif. Pendekatan rule-based sederhana memang bisa

digunakan sebagai filter awal, tetapi mudah terkecoh. Metode machine learning klasik memberikan hasil lebih baik, namun masih terbatas pada pola kata. Pendekatan deep learning, khususnya Transformer, saat ini menjadi tulang punggung sistem deteksi karena mampu memahami konteks bahasa yang kompleks dan dinamis di media sosial.

1. Tantangan dalam Deteksi Ujaran Kebencian

Mendeteksi ujaran kebencian bukanlah hal yang mudah karena beberapa alasan:

a) Perbedaan antara Kritik dan Kebencian
   Kritik sah sering kali dikemas dalam bahasa keras, namun tidak selalu masuk kategori kebencian. Misalnya: *"Pemerintah gagal mengelola ekonomi"* adalah kritik, sedangkan *"Pemerintah tolol karena agama mayoritasnya ini"* adalah ujaran kebencian. Perbedaan tipis ini menuntut sistem dapat memahami konteks, bukan sekadar mendeteksi kata kasar.
b) Penyamaran dalam Slang, Humor, dan Sarkasme
   Banyak pengguna menyamarkan ujaran kebencian dalam bentuk guyonan, meme, atau penggunaan bahasa gaul. Misalnya kata "anjay" bisa digunakan sebagai ekspresi kagum, tapi dalam konteks lain bisa bernuansa menghina. Begitu pula sarkasme: *"Wah, pinter banget dia, bikin hoax tiap hari"* terdengar seperti pujian, tapi sebenarnya sindiran negatif.
c) Penggunaan *Code Words* atau Istilah Kode
   Ujaran kebencian sering tidak disampaikan secara eksplisit, melainkan melalui istilah khusus yang hanya dipahami komunitas tertentu. Contoh di Indonesia:
   - "Cina ijo" → sindiran bernuansa rasial.
   - "Kadal gurun" → istilah peyoratif untuk kelompok tertentu.
   - "Bunglon" → digunakan untuk menstigma kelompok politik lawan. Sistem moderasi perlu mengenali dinamika istilah ini, yang terus berubah dari waktu ke waktu.

2. Pendekatan NLP dalam Deteksi Ujaran Kebencian

a) Rule-Based (Pendekatan Awal). Pendekatan ini menggunakan daftar kata kasar atau larangan tertentu untuk memfilter konten. Misalnya, jika kalimat mengandung kata makian, maka otomatis ditandai sebagai ujaran kebencian.
    1) Contoh: teks *"Dasar bodoh, kamu kafir!"* akan ditandai karena memuat kata kasar + unsur diskriminasi agama.
    2) Kelemahan: tidak mampu membedakan konteks. Kalimat *"Dia bilang saya bodoh, tapi saya ketawa saja"* bisa salah ditandai padahal bukan ujaran kebencian.

b) Machine Learning Klasik (Naive Bayes, SVM, Random Forest). Metode ini menganalisis teks berdasarkan fitur sederhana, seperti frekuensi kata, n-gram, atau skor TF-IDF. Dengan dataset yang cukup, algoritma bisa belajar pola umum dari ujaran kebencian.
    1) Contoh: teks dengan kombinasi kata "agama X + hinaan" akan diklasifikasi sebagai ujaran kebencian.
    2) Kelebihan: cukup efektif untuk teks pendek dan data terstruktur.
    3) Keterbatasan: gagal menangkap makna sarkasme atau metafora, serta kurang adaptif terhadap slang baru.

c) Deep Learning (CNN, RNN, LSTM, Transformer). Pendekatan ini menggunakan jaringan saraf tiruan yang mampu memahami konteks kalimat secara menyeluruh. Model berbasis Transformer seperti BERT atau IndoBERT dapat mengenali makna ujaran kebencian meskipun kata-kata yang dipakai disamarkan.
    1) Contoh: kalimat *"Dia hitam tapi pintar banget"* bisa dikenali sebagai diskriminasi karena model paham bahwa kata "hitam" digunakan dengan nada stereotipe.
    2) Kelebihan: mampu menangkap konteks, slang, bahkan *code-switching*.
    3) Keterbatasan: membutuhkan data pelatihan yang sangat besar dan sumber daya komputasi tinggi.

3. Contoh Studi Kasus: Deteksi Ujaran Kebencian di Twitter Indonesia

Misalnya kita menggunakan data yang bersumber dari Twitter (atau X) merupakan salah satu media sosial paling populer di Indonesia. Platform ini sering digunakan untuk menyampaikan opini politik, sosial, maupun budaya. Namun, karakter teks yang singkat (280 karakter), penuh slang, singkatan, dan campur bahasa, membuat Twitter juga menjadi ruang subur bagi penyebaran ujaran kebencian, terutama menjelang momen politik atau isu sensitif agama dan ras.

a) Dataset

- Sumber: Ribuan tweet publik dalam bahasa Indonesia dikumpulkan menggunakan Twitter API.
- Labeling: Tweet dikategorikan ke dalam tiga kelas:
    1. Ujaran kebencian (hate speech) – menyerang kelompok identitas (ras, agama, gender, dll.).
    2. Abusive language (kasar/bullying) – penghinaan personal tanpa menyerang identitas kelompok.
    3. Netral – opini atau komentar tanpa unsur kebencian.
- Contoh data:
    1. Hate speech: *"Dasar kafir, agama lo sesat!"*
    2. Abusive: *"Kamu goblok banget sih main game aja kalah."*
    3. Netral: *"Saya suka makan bakso di warung dekat rumah."*

b) Metode

1. Rule-Based Filter → digunakan sebagai tahap awal untuk menyaring kata kasar umum.
2. Machine Learning (SVM, Naive Bayes) → dilatih dengan fitur TF-IDF untuk mengenali pola ujaran kebencian.
3. Deep Learning (IndoBERT) → digunakan untuk memahami konteks penuh tweet, termasuk slang, sarkasme, dan code-switching.

c) Implementasi Deteksi Ujaran Kebencian (End-to-End Pipeline)

1. Load Dataset

Dataset biasanya berupa file CSV dengan dua kolom utama: teks dan label (misalnya: *hate_speech*, *abusive*, *neutral*).

```
import pandas as pd

# Load dataset
data = pd.read_csv("tweet_dataset.csv")
print(data.head())
```

Contoh isi dataset:

Teks	Label
"Dasar kafir, agama lo sesat!"	hate_speech
"Kamu goblok banget sih main game aja kalah."	abusive
"Saya suka makan bakso di warung dekat rumah."	neutral

2. Preprocessing

Langkah ini menyiapkan teks agar bisa diproses model:

- Lowercasing (huruf kecil semua).
- Menghapus URL, mention, hashtag.
- Tokenisasi kata.
- Menghapus stopwords.

```
import re
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')
stop_words = set(stopwords.words('indonesian'))

def preprocess(text):
    text = text.lower()
    text = re.sub(r"http\S+|www\S+|https\S+", '', text)  # hapus URL
    text = re.sub(r"@\w+|#\w+", '', text)                # hapus mention/hashtag
    tokens = text.split()
    tokens = [t for t in tokens if t not in stop_words]  # hapus stopwords
    return " ".join(tokens)

data["clean_text"] = data["Teks"].apply(preprocess)
print(data.head())
```

Contoh hasil preprocessing:

⇒ Input: *"Join room 88 dapet chip gratis bro, legit banget!"*

⇒ Output: *"join room 88 dapet chip gratis bro legit banget"*

3. Split Dataset (Training & Testing)

Dataset dibagi menjadi data latih dan uji.

```
from sklearn.model_selection import train_test_split

X = data["clean_text"]
y = data["Label"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4. Feature Extraction (TF-IDF)

Untuk Naive Bayes / SVM, teks diubah menjadi representasi numerik dengan TF-IDF.

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(ngram_range=(1,2))  # unigram + bigram
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

5. Modeling

a) Naive Bayes

```
from sklearn.naive_bayes import MultinomialNB

nb_model = MultinomialNB()
nb_model.fit(X_train_tfidf, y_train)

y_pred_nb = nb_model.predict(X_test_tfidf)
```

b) SVM

```
from sklearn.svm import LinearSVC

svm_model = LinearSVC()
svm_model.fit(X_train_tfidf, y_train)

y_pred_svm = svm_model.predict(X_test_tfidf)
```

c) IndoBERT (Deep Learning)

```
from transformers import BertTokenizer,
BertForSequenceClassification
from transformers import Trainer, TrainingArguments

tokenizer =
BertTokenizer.from_pretrained("indobenchmark/indobert-
base-p1")
model =
BertForSequenceClassification.from_pretrained("indobenchm
ark/indobert-base-p1", num_labels=3)

# Tokenisasi dataset
```

```
train_encodings = tokenizer(list(X_train),
truncation=True, padding=True, max_length=128)
test_encodings = tokenizer(list(X_test), truncation=True,
padding=True, max_length=128)

import torch
class HateDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels
    def __getitem__(self, idx):
        item = {key: torch.tensor(val[idx]) for key, val
in self.encodings.items()}
        item['labels'] =
torch.tensor(int(self.labels.iloc[idx]))
        return item
    def __len__(self):
        return len(self.labels)

train_dataset = HateDataset(train_encodings, y_train)
test_dataset = HateDataset(test_encodings, y_test)

# Training
training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="epoch",
    num_train_epochs=2,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    logging_dir="./logs",
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=test_dataset
)

trainer.train()
results = trainer.evaluate()
print(results)
```

6. Evaluasi Model

Menggunakan akurasi, precision, recall, dan F1-score.

```
from sklearn.metrics import classification_report

print("Naive Bayes:\n", classification_report(y_test, y_pred_nb))
print("SVM:\n", classification_report(y_test, y_pred_svm))
```

Hasil Eksperimen (simulasi saja):

Model	Akurasi	Precision	Recall	F1-score
Naive Bayes	72%	0.70	0.68	0.69
SVM	78%	0.76	0.75	0.75
IndoBERT	88%	0.87	0.86	0.86

6.2 Deteksi Hoaks dan Misinformasi

Hoaks dan misinformasi adalah konten teks yang berisi klaim palsu, menyesatkan, atau belum terverifikasi. Bedanya:

- Hoaks → informasi salah yang dibuat dengan sengaja untuk menipu atau memanipulasi opini publik.
- Misinformasi → informasi salah yang disebarkan tanpa niat menipu, biasanya karena ketidaktahuan.

Deteksi hoaks dan misinformasi memerlukan pendekatan multi-layered: filter kata emosional, verifikasi fakta otomatis, serta model NLP cerdas untuk memahami klaim kompleks. Di Indonesia, kolaborasi dengan situs cek fakta resmi (seperti Kominfo dan Mafindo) sangat penting agar sistem AI tetap terhubung dengan sumber data yang terpercaya.

Di era digital, hoaks dan misinformasi sering menyebar cepat karena memanfaatkan sifat viral media sosial. Isu politik, kesehatan, dan bencana alam menjadi target utama penyebaran, misalnya hoaks tentang vaksin COVID-19 atau berita palsu terkait hasil pemilu.

1. Tantangan dalam Deteksi

a) Bahasa Persuasif dan Emosional. Banyak hoaks menggunakan kata-kata hiperbolis seperti “TERBUKTI 100%!!!” atau “WAJIB SEBARKAN!”. Kalimat seperti ini menarik emosi pembaca, membuat mereka lebih mudah percaya.
b) Kemiripan dengan Berita Asli. Beberapa hoaks disusun dengan gaya jurnalistik, lengkap dengan judul, tanggal, dan sumber palsu, sehingga sulit dibedakan dari berita sahih.
c) Fakta Parsial. Sebagian besar hoaks tidak sepenuhnya salah, melainkan mencampur fakta benar dengan informasi menyesatkan. Ini membuat model deteksi harus lebih cerdas dalam membedakan.

d) Bahasa Gaul dan Code-Switching. Sama seperti ujaran kebencian, hoaks di media sosial sering menggunakan campuran bahasa atau slang agar lebih natural.

2. Pendekatan NLP dalam Deteksi Hoaks

a) Text Classification (Supervised Learning)
   a. Teks berita atau postingan diproses menjadi fitur (TF-IDF, word embeddings).
   b. Dilatih dengan algoritma ML (SVM, Random Forest, CNN, Transformer).
   c. Contoh: berita dengan banyak kata emosional → lebih cenderung dikategorikan hoaks.

b) Fact-Checking Otomatis
   a. Klaim dalam teks dibandingkan dengan basis data fakta (misalnya WHO, Kominfo, atau situs cek fakta resmi).
   b. Contoh: klaim *"Vaksin COVID-19 menyebabkan kematian massal"* → dicek ke WHO, hasilnya *false*.

c) Transformer Models (BERT, IndoBERT, RoBERTa)
   a. Mampu memahami konteks klaim, membedakan opini vs fakta.
   b. Cocok untuk mendeteksi hoaks kompleks yang tidak hanya bergantung pada kata kunci.

Contoh Kasus Teks Hoaks: *"BREAKING NEWS! Pemerintah sudah diam-diam melegalkan judi online, makanya banyak situs baru muncul. SEBARKAN sebelum dihapus!"*

   a. Rule-based → hanya melihat kata "judi online" → belum tentu hoaks.
   b. ML klasik → bisa salah karena banyak kata umum.
   c. Transformer → bisa mengenali pola "breaking news + sebarkan + klaim palsu" → hoaks.

3. Contoh Implementasi Deteksi Hoaks dan Misinformasi

a. Load Dataset

Dataset biasanya berisi artikel berita, posting media sosial, atau klaim publik yang sudah diberi label (*hoax* atau *valid*).

```
import pandas as pd

# Contoh dataset berita
data = pd.read_csv("news_dataset.csv")
print(data.head())
```

Contoh dataset:

Judul	**Isi Berita**	**Label**
"Vaksin COVID-19 menyebabkan kematian massal"	Klaim palsu yang tidak terbukti secara ilmiah.	Hoax
"Gempa 6,5 SR terjadi di Maluku Utara"	Fakta, diverifikasi oleh BMKG.	Valid

b. Preprocessing

Teks diproses agar bersih dan konsisten:

- Lowercase.
- Hapus angka, URL, simbol.
- Tokenisasi.
- Opsional: stemming/lemmatization.

```
import re

def preprocess(text):
    text = text.lower()
    text = re.sub(r"http\S+|www\S+", '', text)  # hapus URL
    text = re.sub(r"[^a-zA-Z\s]", '', text)     # hapus simbol/angka
    return text

data["clean_text"] = data["Isi Berita"].apply(preprocess)
```

c. Split Dataset

```
from sklearn.model_selection import train_test_split

X = data["clean_text"]
y = data["Label"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

d. Feature Extraction

Mengubah teks menjadi vektor numerik dengan TF-IDF.

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(max_features=5000, ngram_range=(1,2))
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

e. Modeling

1) SVM

```
from sklearn.svm import LinearSVC
from sklearn.metrics import classification_report

svm_model = LinearSVC()
svm_model.fit(X_train_tfidf, y_train)

y_pred_svm = svm_model.predict(X_test_tfidf)
print("SVM:\n", classification_report(y_test, y_pred_svm))
```

b) Random Forest

```
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(n_estimators=200, random_state=42)
rf_model.fit(X_train_tfidf, y_train)

y_pred_rf = rf_model.predict(X_test_tfidf)
print("Random Forest:\n", classification_report(y_test, y_pred_rf))
```

c) IndoBERT (Transformer)

```
from transformers import BertTokenizer,
BertForSequenceClassification, Trainer, TrainingArguments

tokenizer =
BertTokenizer.from_pretrained("indobenchmark/indobert-
base-p1")
model =
BertForSequenceClassification.from_pretrained("indobenchm
ark/indobert-base-p1", num_labels=2)

train_encodings = tokenizer(list(X_train),
truncation=True, padding=True, max_length=128)
test_encodings = tokenizer(list(X_test), truncation=True,
padding=True, max_length=128)

import torch
class NewsDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):
        self.encodings = encodings
        self.labels = labels
    def __getitem__(self, idx):
        item = {key: torch.tensor(val[idx]) for key, val
in self.encodings.items()}
```

```
        item['labels'] =
torch.tensor(int(self.labels.iloc[idx] == "Hoax"))
        return item
    def __len__(self):
        return len(self.labels)

train_dataset = NewsDataset(train_encodings, y_train)
test_dataset = NewsDataset(test_encodings, y_test)

training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="epoch",
    num_train_epochs=2,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    logging_dir="./logs",
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=test_dataset
)

trainer.train()
results = trainer.evaluate()
print(results)
```

f. Evaluasi

Gunakan akurasi, precision, recall, F1-score.

Hasil simulasi:

Model	**Akurasi**	**Precision**	**Recall**	**F1-score**
SVM	80%	0.78	0.77	0.77
Random Forest	82%	0.80	0.79	0.79
IndoBERT	90%	0.89	0.88	0.88

Deteksi hoaks dan misinformasi merupakan tantangan penting dalam moderasi konten teks karena sifatnya yang sering menyerupai berita asli, penuh bahasa persuasif, serta bercampur antara fakta dan opini. Algoritma klasik seperti Support Vector Machine (SVM) dan Random Forest (RF) terbukti cukup efektif sebagai baseline, terutama ketika dipadukan dengan representasi teks berbasis TF-IDF. Model ini mampu mengenali pola umum dalam teks, seperti penggunaan kata-

kata hiperbolis atau struktur kalimat tertentu yang sering muncul pada hoaks. Namun demikian, untuk klaim yang lebih kompleks, ambigu, atau mengandung konteks sosial tertentu, model klasik seringkali kurang memadai. Di sinilah IndoBERT dan model berbasis Transformer lainnya menunjukkan keunggulannya. Dengan kemampuan memahami konteks semantik secara mendalam, IndoBERT dapat mendeteksi hoaks yang terselubung di balik bahasa yang tampak wajar, termasuk teks yang menggunakan slang atau campuran bahasa (*code-switching*).

Ke depan, deteksi hoaks akan semakin kuat apabila sistem moderasi konten tidak hanya mengandalkan NLP, tetapi juga diintegrasikan dengan basis data fact-checking resmi. Misalnya, klaim yang ditemukan dapat secara otomatis diverifikasi dengan database Kominfo, Mafindo, atau lembaga cek fakta global. Dengan kombinasi ini, sistem moderasi konten digital akan mampu menghadirkan pendekatan yang lebih akurat, adaptif, dan relevan dengan dinamika informasi di media sosial Indonesia.

6.3 Deteksi Promosi Ilegal (Narkoba, Judi, dsb.)

Promosi ilegal di media sosial adalah segala bentuk teks yang berisi ajakan, iklan, atau tautan terkait aktivitas yang dilarang hukum, seperti perjudian online, penjualan narkoba, pornografi, hingga penipuan finansial. Promosi ini seringkali dikemas dengan bahasa persuasif, hadiah gratis, atau disamarkan dalam bentuk slang dan simbol agar lolos dari sistem deteksi otomatis.

Di Indonesia, judi online menjadi salah satu masalah serius karena maraknya iklan terselubung di Facebook, Instagram, WhatsApp, dan TikTok. Contoh paling umum adalah kalimat promosi dengan janji “chip gratis”, “bonus saldo”, atau “room 88” yang disebarkan secara masif. Hal serupa juga berlaku untuk promosi narkoba melalui istilah kode seperti “barang enak”, “paket happy”, atau “obat stamina”.

Deteksi promosi ilegal di media sosial membutuhkan pendekatan multi-layered:

- Rule-based/Regex untuk mendeteksi pola sederhana.

- Machine Learning/Deep Learning untuk menangkap konteks dan variasi bahasa.
- OCR + NLP untuk mendeteksi promosi terselubung dalam gambar/video.

Dengan kombinasi teknik ini, sistem moderasi dapat lebih efektif menyaring iklan judi, narkoba, atau konten berbahaya lainnya yang marak di platform digital.

1. Tantangan dalam Deteksi Promosi Ilegal
    a. Penyamaran Kata Kunci. Pengiklan sering menggunakan variasi huruf dan angka agar lolos filter. Contoh: *"sl0t 88"* (slot 88), *"jvdi"* (judi), *"d4pat chip gr@tis"*.
    b. Kalimat Persuasif dan Netralisasi Bahasa. Promosi sering dikemas dengan gaya bahasa santai agar tidak mencurigakan. Contoh: *"Gabung grup WA, dapat hadiah saldo tiap malam guys!"*.
    c. Konten Campuran dengan Gambar/Video. Banyak iklan judi atau narkoba muncul dalam bentuk poster visual atau meme, dengan teks overlay yang hanya bisa dibaca setelah diekstraksi menggunakan OCR.
    d. Evolusi Cepat Pola Bahasa. Bahasa iklan ilegal terus berubah mengikuti tren agar lolos moderasi. Kata-kata baru, singkatan, atau kode tertentu muncul hampir setiap minggu.

2. Pendekatan NLP untuk Deteksi Promosi Ilegal
    a. Rule-Based dan Regex. Digunakan untuk mendeteksi pola sederhana seperti nomor WhatsApp, link website, atau kata kunci spesifik. Contoh: regex untuk mendeteksi "+62xxxx" atau "bit.ly/slot".
    b. Machine Learning (SVM, Random Forest). Menggunakan fitur TF-IDF dari teks iklan. Cukup efektif untuk membedakan antara teks promosi dan teks percakapan normal.
    c. Deep Learning (CNN, Transformer). Mampu mengenali pola bahasa gaul, variasi huruf, dan konteks ajakan. Contoh: kalimat *"Main slot 88 legit, chip gratis nunggu kamu!"* dapat

dideteksi sebagai iklan judi, meski tidak eksplisit menggunakan kata “judi online”.

d. OCR + NLP (Multimodal). Untuk mendeteksi promosi dalam gambar/banner, teks diekstraksi dengan OCR, lalu dianalisis dengan model NLP. Contoh: poster dengan teks “Bonus 100% untuk member baru” → dikenali sebagai promosi judi.

Contoh Kasus

- Teks Promosi: *“Join room 88 dapet chip gratis bro, legit banget!”*
  - Rule-based → mengenali kata “chip gratis” → tandai sebagai promosi.
  - SVM → mengklasifikasi sebagai iklan karena pola ajakan.
  - Transformer → lebih presisi, karena memahami kombinasi ajakan + kata kunci → promosi judi online.
- Teks Netral: *“Room 88 di hotel itu enak banget buat rapat.”*
  - Rule-based → mungkin salah deteksi karena ada kata “room 88”.
  - Transformer → benar mengklasifikasi sebagai konten netral, karena konteksnya berbeda.

BAB VII
Moderasi Konten Visual (Gambar & Video)

Konten visual dalam bentuk gambar dan video kini mendominasi media sosial seperti Instagram, TikTok, dan YouTube. Jika moderasi teks berfokus pada komentar atau caption, maka moderasi visual harus menangani elemen visual yang jauh lebih kompleks. Tantangan utama adalah konten berbahaya sering kali disamarkan dalam bentuk visual, seperti poster judi online, meme bernada ujaran kebencian, atau video yang mengandung kekerasan eksplisit.

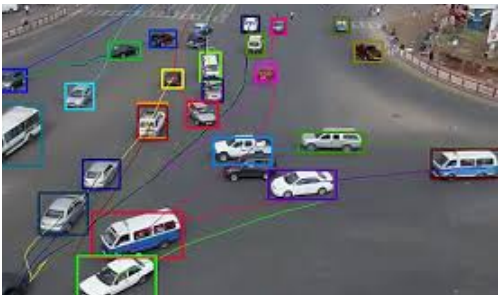
Deteksi otomatis pada konten visual mengandalkan Computer Vision dan integrasi dengan OCR (Optical Character Recognition) serta ASR (Automatic Speech Recognition), terutama untuk konten multimodal.

Bab ini akan membahas tiga aplikasi utama moderasi konten visual:

1. Pengenalan Gambar Eksplisit
2. OCR untuk Teks dalam Gambar/Video
3. Analisis Video untuk Konten Sensitif

7.1 Pengenalan Gambar Eksplisit

Gambar eksplisit merujuk pada konten visual yang menampilkan pornografi, kekerasan ekstrem, atau materi visual yang tidak pantas, yang dapat membahayakan pengguna terutama anak-anak dan remaja. Di media sosial, gambar eksplisit tidak hanya hadir secara frontal (misalnya foto pornografi), tetapi juga dalam bentuk poster promosi judi, ilustrasi kekerasan, atau meme berbau SARA. Pada Gambar 7.1 memberikan ilustrasi deteksi objek citra.



Gambar 7. 1 Deteksi Objek

Bagi platform media sosial, deteksi gambar eksplisit menjadi prioritas utama karena dampaknya sangat besar:

- Psikologis → paparan konten pornografi atau kekerasan dapat merusak kesehatan mental pengguna.
- Sosial → penyebaran gambar diskriminatif dapat memperkuat ujaran kebencian dan intoleransi.
- Hukum → distribusi pornografi anak, narkoba, atau judi online jelas melanggar hukum di banyak negara, termasuk Indonesia.

Deteksi gambar eksplisit merupakan salah satu pilar penting dalam moderasi konten visual. Pendekatan klasik berbasis warna kulit sudah tidak relevan karena tingkat kesalahannya tinggi. Kini, pendekatan modern menggunakan CNN dan Vision Transformer terbukti lebih akurat. Integrasi dengan OCR + NLP membuat sistem semakin kuat dalam mendeteksi promosi ilegal yang sering disamarkan dalam bentuk poster atau meme.

Dengan teknologi ini, platform media sosial dapat secara proaktif memblokir gambar berbahaya sebelum sampai ke pengguna, sekaligus mematuhi regulasi hukum yang berlaku. Berikut diagram alur deteksi gambar eksplisit (Input gambar → CNN/ViT → OCR → Klasifikasi → Output: Normal/Eksplisit/Promosi Ilegal).

Gambar 7. 2 Diagram Alur Deteksi Gambar

1. Input Gambar. Tahap awal adalah menerima gambar dari berbagai sumber, seperti unggahan pengguna, crawling web, atau input API. Gambar biasanya berformat JPEG, PNG, atau HEIC. Pada tahap ini dilakukan validasi dasar, misalnya mengecek ukuran file, tipe format, serta metadata gambar. Hal ini penting untuk memastikan bahwa input yang masuk sesuai dengan standar sistem dan aman diproses lebih lanjut.
2. CNN/ViT (Feature Extraction). Gambar yang sudah tervalidasi kemudian diproses menggunakan model Computer Vision, seperti Convolutional Neural Network (CNN) atau Vision Transformer (ViT). Model ini bertugas mengekstrak ciri visual penting, seperti pola, tekstur, objek, atau bentuk yang ada di dalam gambar. Hasil ekstraksi berupa representasi numerik (embedding) yang memudahkan sistem dalam membedakan antara konten normal, eksplisit, maupun konten promosi ilegal.
3. OCR (Optical Character Recognition). Setelah ciri visual didapatkan, sistem juga melakukan analisis teks pada gambar menggunakan OCR. Tahap ini mendeteksi dan mengenali teks yang muncul di dalam gambar, misalnya nomor telepon, alamat web, atau kata-kata kunci tertentu. Informasi teks ini menjadi penunjang penting, karena banyak konten eksplisit atau promosi ilegal biasanya disisipkan dalam bentuk tulisan pada gambar.
4. Klasifikasi. Hasil dari CNN/ViT dan OCR kemudian digabungkan dan dianalisis oleh modul klasifikasi. Pada tahap ini, model machine learning atau deep learning memutuskan kategori gambar berdasarkan bukti visual dan teks. Tiga kemungkinan hasil klasifikasi adalah: Normal (tidak berbahaya), Eksplisit (mengandung konten seksual/dewasa), dan Promosi Ilegal (misalnya iklan prostitusi atau jual-beli barang terlarang).
5. Output. Tahap akhir adalah memberikan keputusan sistem berdasarkan hasil klasifikasi. Jika gambar termasuk kategori Normal, maka bisa langsung diterima. Jika termasuk Eksplisit atau Promosi Ilegal, maka akan ditandai, diblokir, atau diteruskan ke proses moderasi lebih lanjut sesuai kebijakan. Selain itu,

sistem juga bisa menyimpan log keputusan sebagai bukti dan untuk kebutuhan audit.

7.2 OCR Untuk Teks Dalam Gambar/Video

Optical Character Recognition (OCR) adalah teknologi yang digunakan untuk mengenali teks dalam gambar atau video dan mengubahnya menjadi format digital yang dapat diproses oleh komputer. Dalam konteks moderasi konten, OCR berperan penting untuk mendeteksi pesan tersembunyi di dalam poster, banner, meme, screenshot, hingga subtitle dalam video.

OCR adalah jembatan penting yang menghubungkan analisis visual (Computer Vision) dengan analisis bahasa (NLP). Tanpa OCR, moderasi konten hanya akan efektif untuk gambar eksplisit, tetapi gagal menangkap pesan berbahaya dalam teks visual. Dengan menggabungkan OCR + NLP, sistem moderasi dapat:

- Mendeteksi promosi judi, narkoba, atau penipuan dalam poster/iklan.
- Mengidentifikasi meme bernuansa ujaran kebencian.
- Memproses video dengan teks overlay.

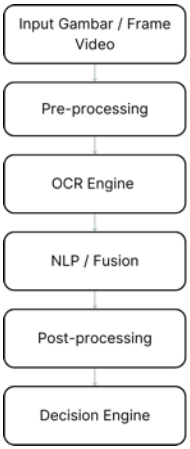
Teknologi OCR yang dikombinasikan dengan model AI modern menjadikan moderasi konten visual lebih akurat, adaptif, dan relevan dengan pola komunikasi digital saat ini.

Mengapa OCR Penting dalam Moderasi Konten?

Banyak promosi ilegal atau ujaran kebencian tidak disebarkan hanya lewat teks biasa, melainkan melalui visual berteks.

- Poster Judi Online: biasanya berisi kata-kata seperti *"bonus 100%", "chip gratis", "slot 88"*.
- Meme Ujaran Kebencian: teks singkat yang menyerang kelompok tertentu ditambahkan ke gambar lucu atau sindiran politik.
- Screenshot Chat Hoaks: digunakan untuk memperkuat klaim palsu.
- Video Promosi Terselubung: teks muncul sebagai overlay atau subtitle singkat yang jika tidak di-OCR akan luput dari deteksi.

Tanpa OCR, sistem moderasi konten hanya bisa mendeteksi gambar secara visual (misalnya nudity atau kekerasan), tetapi gagal menangkap pesan teks berbahaya yang tersembunyi di dalamnya.



Gambar 7. 3 Diagram Alur kerja sistem deteksi iklan judi online berbasis gambar dan video

1) Input Gambar / Frame Video. Tahap awal dimulai dengan menerima masukan berupa gambar atau potongan frame dari video. Sumber data bisa berasal dari unggahan pengguna, crawling media sosial, API, maupun snapshot frame hasil sampling video (misalnya 1 fps, 3 fps, atau berbasis event seperti perubahan scene).
   Pada tahap ini, sistem melakukan validasi dasar, meliputi pengecekan format file (JPEG, PNG, WebP, HEIC), ukuran, orientasi, serta metadata. Untuk data video, kebijakan sampling (sampling policy) sangat penting agar tidak semua frame diproses, tetapi tetap mewakili isi video.
2) Pre-processing (Praproses Gambar). Tujuan tahap ini adalah meningkatkan kualitas sinyal visual agar modul OCR dapat bekerja lebih akurat. Beberapa langkah umum mencakup:
   - Resize & normalisasi sambil menjaga aspek rasio.
   - Denoising untuk mengurangi noise atau artefak kompresi.

- Contrast enhancement (misalnya CLAHE, adaptive histogram equalization).
- Binarization/thresholding untuk teks kontras tinggi.
- Deskew & dewarp untuk memperbaiki teks miring atau foto dokumen yang tidak rata.
- Super-resolution (SRGAN/EDSR) untuk memperbesar teks kecil.
- Konversi color space (ke grayscale atau memanfaatkan channel warna tertentu).
- Cropping heuristics menggunakan object detection agar fokus hanya pada area yang relevan.

Dengan langkah ini, gambar yang masuk akan lebih "bersih" dan mudah dipahami oleh OCR.

3) OCR Engine (Optical Character Recognition)
Tahap ini dibagi menjadi dua bagian utama:

- Text Detection (lokalisasi teks)
  Mendeteksi letak teks dalam gambar, biasanya dalam bentuk bounding box. Algoritma yang umum dipakai adalah MSER, CRAFT, EAST, atau DBNet, yang mendukung teks multi-orientasi.
- Text Recognition (transkripsi teks)
  Mengubah gambar teks menjadi string karakter. Model yang sering digunakan adalah CRNN+CTC, encoder-decoder berbasis attention, atau transformer (misalnya TrOCR).
- Alternatif praktis: Tesseract, EasyOCR, PaddleOCR, atau layanan OCR cloud (Google Vision, AWS Textract).

Output tahap ini berupa teks mentah disertai confidence score untuk tiap token/box.

4) Post-processing & Normalisasi

Hasil teks dari OCR biasanya masih perlu dibersihkan agar lebih bermakna. Proses ini mencakup:

- Spell correction / LM re-ranking menggunakan n-gram LM atau transformer LM.

- Normalization (lowercase, hapus simbol/karakter tidak penting).
- De-obfuscation untuk teks yang dimanipulasi (contoh: “sl0t” → “slot”).
- Entity extraction untuk mendeteksi nomor telepon, URL, email, atau wallet address.
- Language detection & transliteration bila teks menggunakan campuran bahasa/alfabet.
- Confidence handling & temporal smoothing untuk menggabungkan hasil dari beberapa frame video.

5) NLP / Multimodal Fusion
Tahap ini berfungsi memahami konteks teks yang sudah diekstrak. Proses meliputi:

- Tokenisasi & embedding dengan model seperti IndoBERT/mBERT.
- Klasifikasi teks ke label seperti: *promosi_judi*, *penipuan*, *hate_speech*, *normal*, dll.
- Fusi multimodal, yaitu menggabungkan bukti dari OCR, metadata, logo/objek visual, atau transkrip audio (ASR). Misalnya, teks “slot88” + logo kasino = confidence tinggi untuk promosi judi.
- Rule-based overrides untuk pola tertentu (misalnya “deposit + nomor rekening” → langsung dikategorikan promosi).

6) Decision Engine (Pengambilan Keputusan)
Tahap akhir adalah menentukan tindakan berdasarkan hasil analisis. Kebijakan biasanya berbasis skor confidence, tingkat keparahan, serta aturan sistem. Contoh kebijakan:

- Auto-block / auto-remove → jika confidence tinggi & konten berbahaya.
- Auto-flag / reduce distribution → jika confidence sedang.
- Human review → jika confidence rendah atau borderline.
- No action → jika konten aman/normal.

Selain itu, sistem juga menyimpan log & evidence berupa teks hasil OCR, bounding box, confidence score, serta timestamp. Hal ini penting untuk audit, banding, maupun pelatihan ulang model di masa depan.

Contoh Kasus

Kasus 1 – Poster Judi Online

- Input: gambar dengan teks *"Join Slot88, Bonus 100% untuk member baru"*.
- OCR → mengekstrak teks "Join Slot88, Bonus 100% untuk member baru".
- NLP → mendeteksi kata "Slot88" + "Bonus" → promosi ilegal.
- Output → Konten Diblokir.

Kasus 2 – Meme Ujaran Kebencian

- Input: meme dengan teks *"Kadal gurun jangan dipercaya"*.
- OCR → mengekstrak teks tersebut.
- NLP → model Transformer mendeteksi ujaran kebencian terselubung.
- Output → Konten Ditandai sebagai Hate Speech.

Kasus 3 – Video dengan Overlay Teks

Input: video TikTok dengan teks overlay *"Dapetin chip gratis di link ini"*.

- OCR pada frame video → mengekstrak teks.
- NLP → klasifikasi sebagai iklan judi.
- Output → Konten Ilegal.

Tantangan dalam OCR untuk Moderasi

- Variasi Font dan Warna → teks sering dibuat berwarna-warni agar sulit terbaca mesin.
- Teks Berbaur dengan Background → misalnya tulisan kecil di atas gambar kompleks.
- Bahasa Gaul & Ejaan Kreatif → contoh: *"sl0t g@cha gratis"*.

- Noise dan Kualitas Rendah → pada video TikTok atau screenshot WhatsApp.
- Campuran Bahasa (code-switching) → *"Bro, ini 100% legit slot88, no scam!"*.

7.3 Analisis Video Untuk Konten Sensitif

Analisis video untuk moderasi konten sensitif adalah proses mendeteksi, mengidentifikasi, dan mengklasifikasikan unsur berbahaya atau tidak pantas yang muncul dalam video digital. Unsur ini bisa berupa:

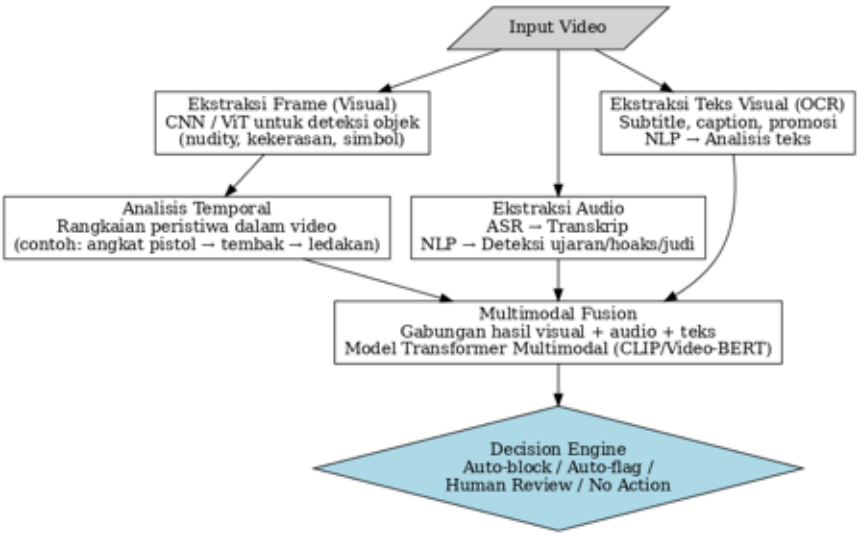
- Kekerasan eksplisit (perkelahian, darah, senjata).
- Pornografi atau konten seksual.
- Promosi ilegal (judi online, narkoba) dalam bentuk overlay teks, audio, atau visual.
- Ujaran kebencian dalam bentuk kombinasi teks + suara + gambar.

Video termasuk konten multimodal, artinya terdiri atas visual (frame/gambar), audio, dan teks yang muncul sebagai subtitle/overlay. Hal ini membuat analisis video lebih kompleks dibanding moderasi teks atau gambar saja.

Mengapa Moderasi Video Penting?

- Dominasi konten digital → TikTok, YouTube, Instagram Reels, dan Shorts adalah bentuk utama interaksi di media sosial.
- Kecepatan penyebaran → sebuah video viral dapat mencapai jutaan penonton dalam hitungan jam.
- Potensi bahaya → video hoaks, propaganda, atau promosi judi online sering dikemas secara menghibur sehingga sulit dideteksi secara manual.

Alur Analisis Video untuk Moderasi Konten



Gambar 7. 4 Diagram Alur Moderasi Konten Multimodal dengan Integrasi Analisis Visual, Teks, dan Audio untuk Pengambilan Keputusan Otomatis

1) Ekstraksi Frame (Analisis Visual). Tahap awal adalah memecah video menjadi potongan gambar (*frame extraction*). Umumnya digunakan kebijakan sampling seperti 1 frame per detik, 3 frame per detik, atau berbasis event (misalnya perubahan scene). Hal ini bertujuan agar sistem tidak perlu menganalisis seluruh frame, melainkan hanya representasi yang cukup menggambarkan isi video. Setiap frame yang diambil kemudian dianalisis dengan model Computer Vision, seperti Convolutional Neural Network (CNN) atau Vision Transformer (ViT). Model ini mendeteksi objek, pola, atau simbol yang mengindikasikan nudity, kekerasan, penggunaan narkoba, maupun logo/simbol berbahaya. Hasil analisis berupa *feature embedding* dan label klasifikasi awal dari sisi visual.
2) Ekstraksi Audio (Speech Analysis). Selain gambar, konten audio dalam video juga dianalisis. Proses dimulai dengan Automatic Speech Recognition (ASR) untuk mengubah suara menjadi teks transkrip. Selanjutnya, transkrip diproses oleh modul Natural

Language Processing (NLP) untuk mengidentifikasi konten berbahaya, misalnya:

- ujaran kebencian,
- promosi judi atau penipuan,
- konten hoaks atau disinformasi.

  Hasil tahap ini berupa label teks + confidence score yang menunjukkan seberapa kuat indikasi pelanggaran dari audio.

3) Ekstraksi Teks Visual (OCR). Video sering memuat teks tambahan, baik berupa subtitle, caption, watermark, maupun overlay promosi tersembunyi. Tahap ini menggunakan Optical Character Recognition (OCR) untuk mengekstrak teks dari setiap frame video yang relevan. Output berupa string teks beserta lokasi bounding box. Kemudian, modul NLP kembali menganalisis isi teks untuk mendeteksi indikasi iklan ilegal, URL mencurigakan, nomor rekening, atau bahasa yang mengandung ujaran kebencian.
4) Analisis Temporal (Konteks Waktu). Video berbeda dengan gambar tunggal, karena memiliki alur peristiwa yang berlangsung dalam waktu tertentu. Sistem moderasi harus mengenali urutan kejadian (temporal context) agar keputusan lebih akurat.

   Contoh: jika satu frame menunjukkan seseorang mengangkat pistol, frame berikutnya memperlihatkan tembakan, dan frame selanjutnya ledakan, maka gabungan peristiwa ini jelas dikategorikan sebagai konten kekerasan.

   Untuk itu, model analisis video biasanya dilengkapi dengan mekanisme sequence modeling seperti RNN, Temporal CNN, atau Transformer berbasis video.
5) Multimodal Fusion (Penggabungan Modalitas). Hasil dari tiga jalur analisis — visual (CNN/ViT), audio (ASR+NLP), dan teks visual (OCR+NLP) — kemudian digabung menggunakan teknik multimodal fusion. Pendekatan modern menggunakan Transformer multimodal (contoh: CLIP, Video-BERT, atau model multimodal-to-text) yang dapat memahami hubungan antar-modalitas. Dengan fusi ini, sistem mampu mengidentifikasi kasus kompleks, misalnya: teks overlay “slot88” + promosi suara ajakan

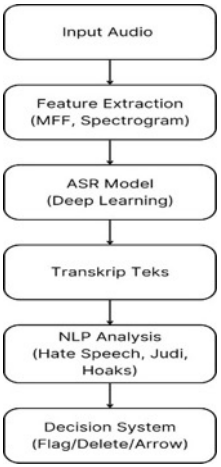
deposit + logo kasino dalam frame → confidence tinggi untuk kategori promosi judi.

6) Keputusan & Moderasi (Decision Engine). Tahap akhir adalah menentukan tindakan terhadap konten video berdasarkan hasil klasifikasi multimodal. Tindakan bersifat policy-driven, dengan beberapa tingkatan:
   - Auto-block / auto-remove → jika confidence tinggi dan pelanggaran jelas (misalnya promosi judi, kekerasan ekstrem, pornografi).
   - Auto-flag / reduce distribution → jika confidence sedang atau masih ada keraguan.
   - Eskalasi ke moderator manusia → untuk kasus ambigu atau borderline, agar manusia memutuskan.
   - No action → jika konten dinilai aman.

BAB VIII
MODERASI KONTEN AUDIO DAN MULTIMODAL

8.1 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) adalah teknologi kecerdasan buatan yang berfungsi untuk mengubah ucapan manusia dalam bentuk audio menjadi teks digital yang dapat diproses lebih lanjut. Dalam konteks moderasi konten digital, ASR memegang peran penting karena banyak konten berbahaya di media sosial disampaikan bukan hanya melalui teks tertulis, tetapi juga dalam bentuk suara pada video, siaran langsung (live streaming), maupun podcast. Dengan ASR, sistem dapat menyalin percakapan atau narasi dari audio menjadi transkrip teks yang kemudian dianalisis menggunakan Natural Language Processing (NLP) untuk mendeteksi ujaran kebencian, promosi judi, penipuan, hingga penyebaran hoaks



Gambar 8. 1 Alur Proses Automatic Speech Recognition

Keunggulan utama ASR dalam moderasi konten adalah kemampuannya memperluas jangkauan deteksi ke medium audio. Misalnya, seorang streamer dapat menyamarkan iklan judi dengan menyelipkan kata-kata promosi secara lisan tanpa menuliskannya di

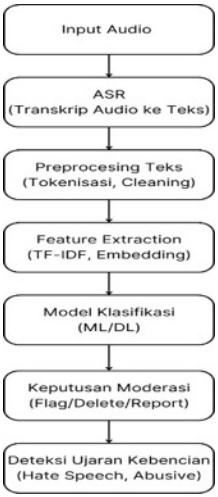
caption atau deskripsi video. Jika hanya mengandalkan analisis teks tertulis, sistem mungkin gagal mengenali pelanggaran tersebut. Namun, dengan ASR, ucapan tersebut dapat ditranskripsi menjadi teks, lalu diklasifikasikan secara otomatis sehingga promosi ilegal dapat terdeteksi. Platform besar seperti YouTube dan TikTok sudah memanfaatkan ASR, baik untuk pembuatan subtitle otomatis maupun sebagai bagian dari sistem deteksi kata-kata sensitif dalam audio.

Meskipun demikian, implementasi ASR juga menghadapi sejumlah tantangan. Faktor seperti kualitas audio yang buruk, adanya musik latar, aksen beragam, penggunaan bahasa campuran (code-switching), serta slang atau istilah gaul dapat menurunkan akurasi transkripsi. Kesalahan transkripsi ini berisiko membuat sistem gagal mengidentifikasi konten berbahaya atau justru menghasilkan false positive. Oleh karena itu, sistem moderasi modern sering mengombinasikan ASR dengan teknik lain seperti Computer Vision (CV) dan Optical Character Recognition (OCR) dalam pendekatan multimodal, sehingga analisis konten tidak hanya bertumpu pada satu jalur, tetapi melibatkan teks, gambar, dan suara sekaligus. Dengan cara ini, ASR menjadi fondasi penting untuk memastikan bahwa konten audio berbahaya juga bisa terdeteksi secara otomatis dan menyeluruh dalam ekosistem media sosial.

8.2 Deteksi Ujaran Toksik Dalam Audio

Deteksi ujaran toksik dalam audio merupakan langkah lanjutan setelah proses Automatic Speech Recognition (ASR). Setelah audio ditranskripsi menjadi teks, sistem moderasi perlu mengidentifikasi apakah di dalamnya terdapat ujaran kebencian, bahasa kasar, atau bentuk kekerasan verbal lainnya yang berpotensi melanggar kebijakan platform. Ujaran toksik mencakup beberapa kategori, seperti *hate speech* (ujaran kebencian berbasis SARA atau identitas), *abusive language* (penghinaan personal, bullying), hingga *offensive content* (komentar merendahkan atau memprovokasi). Dengan deteksi ini, platform media sosial dapat mencegah penyebaran ujaran

kebencian yang sering kali berdampak buruk pada kesehatan mental pengguna dan stabilitas social.



Gambar 8. 2 Contoh Alur Proses Deteksi

Secara teknis, deteksi ujaran toksik dalam audio melibatkan beberapa pendekatan. Pada tahap awal, sistem dapat menggunakan rule-based filter untuk mendeteksi kata-kata kasar yang umum, misalnya melalui daftar kata sensitif (*bad words list*). Namun, pendekatan ini sering tidak cukup karena tidak mempertimbangkan konteks. Oleh karena itu, metode machine learning klasik seperti *Naive Bayes* dan *Support Vector Machine (SVM)* dengan fitur TF-IDF banyak digunakan untuk membedakan antara ujaran normal dan ujaran toksik. Untuk konteks yang lebih kompleks, teknologi deep learning, terutama model berbasis Transformer seperti BERT atau IndoBERT, lebih unggul karena mampu memahami bahasa gaul, sarkasme, hingga campuran bahasa (code-switching) yang sering digunakan di media sosial.

Pipeline deteksi ujaran toksik biasanya terdiri dari beberapa tahap:

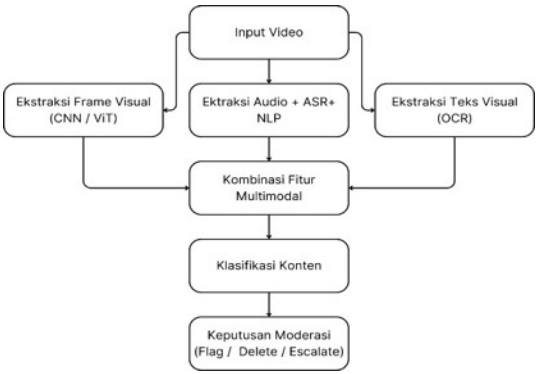
1. Preprocessing Audio → suara direkam dan diubah menjadi teks melalui ASR.

2. Tokenisasi & Ekstraksi Fitur → teks hasil transkripsi diproses dengan metode TF-IDF, word embeddings, atau contextual embeddings dari model Transformer.
3. Klasifikasi Toksisitas → model machine learning atau deep learning menentukan apakah teks termasuk *hate speech*, *abusive*, atau netral.
4. Keputusan Moderasi → jika terdeteksi ujaran toksik, sistem dapat menandai, membatasi distribusi, atau menghapus konten, serta dalam kasus berat, melaporkan pengguna ke moderator manusia.

Tantangan dalam deteksi ujaran toksik meliputi variasi bahasa, penggunaan sarkasme, serta strategi pengguna dalam menyamarkan kata-kata (misalnya mengganti huruf dengan simbol). Oleh karena itu, pendekatan yang efektif biasanya menggabungkan rule-based, machine learning, dan deep learning dalam satu sistem hybrid. Dengan cara ini, platform dapat mencapai keseimbangan antara kecepatan deteksi otomatis dan akurasi kontekstual, sehingga ujaran berbahaya dalam konten audio dapat ditangani secara lebih menyeluruh dan tepat sasaran.

8.3 Integrasi Multimodal: Teks + Gambar + Audio

Integrasi multimodal adalah pendekatan moderasi konten digital yang menggabungkan analisis dari berbagai sumber data sekaligus, yaitu teks, gambar, dan audio. Hal ini penting karena pelanggaran kebijakan platform sering kali tidak muncul hanya dalam satu bentuk. Misalnya, sebuah video promosi judi online bisa menyembunyikan pesan dalam teks kecil di gambar, menyampaikan narasi iklan melalui audio, sekaligus menambahkan caption yang menyesatkan. Dengan mengintegrasikan ketiga modalitas ini, sistem moderasi dapat mendeteksi pelanggaran secara lebih menyeluruh dan akurat .



Gambar 8. 3 Proses Integrasi Multimodal

Proses integrasi multimodal biasanya mencakup beberapa tahap utama. Pertama, sistem melakukan ekstraksi frame visual dari video dan menganalisisnya menggunakan model *Computer Vision* seperti CNN atau Vision Transformer (ViT) untuk mendeteksi objek atau simbol berbahaya (contohnya senjata, narkoba, atau logo perjudian). Kedua, ekstraksi audio dilakukan dengan bantuan ASR untuk mengubah ucapan menjadi teks, kemudian dianalisis menggunakan NLP guna mendeteksi ujaran kebencian, promosi ilegal, atau hoaks. Ketiga, OCR (Optical Character Recognition) digunakan untuk membaca teks dalam gambar atau video, seperti poster, meme, atau subtitle. Hasil analisis dari ketiga sumber ini kemudian digabungkan dalam modul klasifikasi multimodal untuk menghasilkan keputusan akhir.

Keunggulan integrasi multimodal adalah kemampuannya melihat konten dari berbagai sudut. Jika satu jalur analisis gagal, jalur lain dapat menjadi pelengkap. Misalnya, meskipun ASR gagal mengenali ucapan karena kualitas audio buruk, OCR tetap bisa menangkap teks promosi di dalam video. Tantangan utama metode ini adalah kebutuhan komputasi yang tinggi, sinkronisasi antar-modalitas, serta risiko bias jika salah satu model lebih dominan. Namun, dengan perkembangan *deep learning* dan model multimodal modern seperti CLIP atau Video Transformer, integrasi ini semakin efisien dan menjadi standar dalam moderasi konten tingkat lanjut

BAB IX
PENERAPAN KECERDASAN BUATAN UNTUK DETEKSI IKLAN JUDI ONLINE DI MEDIA SOSIAL INDONESIA

Meningkatnya iklan perjudian online di media sosial menjadi ancaman bagi keamanan digital, terutama di platform multibahasa seperti Indonesia. Penelitian ini mengusulkan kerangka kerja ekstraksi teks multimodal berbasis deep learning dan AutoML yang menggabungkan teks, OCR dari gambar, dan ASR dari video untuk mendeteksi konten perjudian. Empat pendekatan utama dievaluasi: RNN, CNN, BERT, dan AutoML, dengan RNN menunjukkan hasil terbaik dalam deteksi iklan perjudian, diikuti oleh CNN dan AutoML. Sementara BERT menunjukkan kinerja yang lebih rendah akibat keterbatasan adaptasi pada teks Indonesia yang penuh kebisingan.

Bab ini akan membahas kerangka kerja multimodal yang diusulkan, termasuk metodologi yang digunakan dalam pengumpulan data, preprocessing, dan pengembangan model. Pembahasan juga mencakup hasil eksperimen yang menunjukkan kinerja masing-masing model serta tantangan utama seperti bahasa promosi tersirat, kebisingan dalam OCR/ASR, dan ambiguitas konteks yang mempengaruhi deteksi iklan perjudian.

9.1 Pendahuluan

Pertumbuhan pesat platform media sosial telah mengubah cara orang berinteraksi, bertukar informasi, dan mengakses hiburan. Transformasi digital ini membawa banyak manfaat, namun juga mempermudah peredaran konten ilegal dan berbahaya, termasuk iklan perjudian online. Industri perjudian online global diperkirakan mencapai nilai pasar USD 63,53 miliar pada tahun 2022 dan diproyeksikan terus tumbuh. Di Indonesia, dengan penetrasi internet yang telah melebihi 78% pada tahun 2023 dan penggunaan media sosial yang termasuk yang tertinggi di Asia Tenggara, perjudian online

menjadi masalah sosial dan hukum yang serius. Meskipun dilarang oleh hukum nasional, operator perjudian memanfaatkan anonimitas dan jangkauan luas platform seperti Facebook, X (sebelumnya Twitter), Instagram, dan YouTube untuk mempromosikan layanan perjudian kepada audiens yang lebih luas, termasuk pengguna muda yang rentan.

Risiko terkait iklan perjudian online melampaui pelanggaran hukum, mencakup dampak sosial seperti meningkatkan perilaku adiktif dan kerugian ekonomi, seperti kerugian finansial di kalangan pemain. Selain itu, kejahatan terkait perjudian dan skema penipuan sering kali mengikuti, menjadikan deteksi dan pencegahan konten ini menjadi perhatian publik. Penegak hukum dan moderator platform menghadapi tantangan besar dalam mengidentifikasi konten semacam ini, terutama karena pesan promosi sering kali tersirat, menggunakan kata-kata kode, slang lokal, atau narasi persuasif untuk menghindari sistem deteksi otomatis. Deteksi konten perjudian dalam teks media sosial sangat menantang karena volume tinggi, keragaman linguistik, dan kebisingan dalam postingan online. Pesan sering kali mengandung bahasa informal, singkatan, emoji, dan kode-switching antara bahasa Indonesia dan Inggris, serta kompleksitas morfologis bahasa Indonesia yang semakin menyulitkan klasifikasi yang akurat. Selain itu, promosi perjudian tidak selalu tertanam dalam teks biasa; mereka bisa muncul dalam gambar, video, atau tersembunyi dalam audio yang diucapkan. Keterbatasan dalam penelitian sebelumnya telah meninggalkan celah deteksi, karena sebagian besar sistem otomatis hanya memproses bidang teks mentah dan mengabaikan konten dari modalitas lainnya.

Berdasarkan tinjauan ini, dua celah penelitian utama diidentifikasi: pertama, tidak adanya kerangka kerja multimodal-to-text yang komprehensif untuk media sosial Indonesia yang mengintegrasikan teks asli, teks yang diekstrak menggunakan OCR dari gambar, dan transkripsi ASR dari suara menjadi representasi yang terstandarisasi; kedua, kurangnya evaluasi sistematik antar model CNN, RNN, model berbasis Transformer, dan AutoML yang dilakukan dalam kondisi preprocessing dan pengaturan

hyperparameter yang konsisten. Untuk mengatasi celah ini, penelitian ini mengusulkan kerangka kerja ekstraksi teks multimodal yang mengonsolidasikan konten terkait perjudian dari teks asli, gambar, frame video, dan transkrip audio menjadi representasi teks yang terstandarisasi. Selanjutnya, analisis perbandingan antar model dilakukan untuk mengevaluasi CNN, LSTM, BiGRU, BERT, dan AutoML (AutoGluon) di bawah kondisi yang identik, memungkinkan penilaian yang adil terhadap kekuatan, kelemahan, dan trade-off mereka dalam hal akurasi, efisiensi, dan interpretabilitas.

9.2 Tinjauan Pustaka

9.2.1 Ekstraksi Teks Multimodal dalam Moderasi Konten

Deteksi konten ilegal di media sosial pada umumnya mengandalkan analisis teks asli. Namun, iklan perjudian sering muncul dalam bentuk gambar, video, atau aliran audio, menjadikan pendekatan unimodal tidak mencukupi. Untuk mengatasi hal ini, teknologi Optical Character Recognition (OCR) telah diterapkan untuk mengekstrak teks yang tersembunyi dalam gambar dan frame video. Sebagai contoh, Thapa et al. (2024) menunjukkan bahwa deteksi ujaran kebencian dalam meme memerlukan strategi multimodal yang menggabungkan fitur visual dan tekstual. Meskipun OCR efektif untuk mengungkapkan sinyal teks yang tersembunyi, kinerjanya sering menurun pada input berkualitas rendah, font bergaya, atau latar belakang yang ramai. Pada konten video, ekstraksi teks dari frame memberikan petunjuk tambahan untuk moderasi. Xu et al. (2024) melaporkan bahwa metode video-OCR terbaru meningkatkan pemahaman multimodal dengan menganalisis teks yang tertanam dalam video. Namun, tantangan seperti kabur gerakan dan biaya komputasi tinggi tetap ada.

Untuk audio, sistem moderasi sering menggunakan Automatic Speech Recognition (ASR) untuk mengubah ucapan menjadi teks sebelum klasifikasi. Bell et al. (2024) menunjukkan bahwa pipeline deteksi pidato beracun sering bergantung pada ASR, namun akurasi sangat sensitif terhadap kebisingan, aksen, dan variasi rekaman. Di Indonesia, Adila et al. (2024) menyoroti kekurangan dataset ASR yang

beragam, sementara kode-switching antara bahasa Indonesia dan Inggris terus menurunkan kualitas transkripsi. Secara keseluruhan, sebagian besar penelitian berfokus pada bahasa Inggris dan pengaturan unimodal, meninggalkan celah dalam penelitian bahasa Indonesia. Sampai saat ini, hanya sedikit pekerjaan yang mencoba mengintegrasikan OCR, ekstraksi teks dari frame video, dan ASR dalam satu alur kerja multimodal untuk deteksi konten berbahaya.

9.2.2 Analisis Perbandingan Model dalam Klasifikasi Teks

Deteksi perjudian online di media sosial Indonesia masih dalam tahap awal. Penelitian sebelumnya mencatat bahwa tata bahasa informal, slang, kode-mixing, dan terbatasnya data anotasi membuat klasifikasi menjadi sangat menantang. Sebagian besar studi menggunakan model individual tanpa perbandingan sistematis antar keluarga model. Metode awal mengandalkan fitur buatan tangan seperti bag-of-words dan TF-IDF dengan pengklasifikasi seperti SVM atau Random Forest, yang efektif dalam deteksi iklan perjudian di Twitter, namun memerlukan desain fitur yang luas dan kesulitan dalam generalisasi pada bahasa yang lebih nuansa atau berkembang. Model deep learning mengurangi kebutuhan akan rekayasa fitur manual. CNN efektif untuk mendeteksi pola n-gram lokal, namun terbatas dalam memodelkan ketergantungan jangka panjang. Model berbasis RNN seperti LSTM dan GRU lebih baik dalam menangkap konteks sekuensial, tetapi pelatihan mereka lebih lambat dan sensitif terhadap kebisingan dalam teks informal. Model berbasis Transformer seperti BERT, yang menjadi standar dalam NLP, menunjukkan hasil yang kuat pada varian bahasa Indonesia seperti IndoBERT, namun seringkali kurang performanya pada data Indonesia yang penuh dengan kebisingan, slang, dan kode campuran tanpa pelatihan domain khusus yang luas.

Framework AutoML, seperti AutoGluon, semakin banyak digunakan untuk tugas klasifikasi teks karena kemampuan otomatisasinya dalam pemilihan model dan tuning hyperparameter. Dalam tugas klasifikasi teks, AutoML dapat mencocokkan kinerja model deep learning yang disetel manual sambil mengurangi upaya

pengembangan. Namun, pipeline default-nya sering gagal untuk sepenuhnya menangkap nuansa linguistik dalam bahasa Indonesia yang kaya morfologi dan informal, sehingga adaptasi domain sangat diperlukan.

9.2.3 Celah Penelitian

Berdasarkan tinjauan pustaka yang ada, dua celah penelitian utama dapat diidentifikasi. Pertama, belum ada kerangka ekstraksi multimodal-to-text yang komprehensif untuk media sosial Indonesia yang mengintegrasikan teks asli, OCR, dan ASR menjadi alur kerja yang terpadu. Kedua, kurangnya evaluasi perbandingan sistematis antara model CNN, RNN, berbasis Transformer, dan AutoML yang dilakukan dalam kondisi eksperimen yang konsisten.

Penelitian ini mengatasi celah-celah tersebut dengan mengusulkan sebuah alur kerja ekstraksi multimodal-to-text yang sepenuhnya terintegrasi dan melakukan analisis benchmarking perbandingan model yang luas. Dengan menyelaraskan pengaturan eksperimen di seluruh model, studi ini memastikan perbandingan yang lebih adil dan memberikan wawasan praktis bagi akademisi dan industri. Tabel 9.1 menyajikan ikhtisar dari studi-studi relevan sebelumnya, fokus metodologi, dan temuan kunci, yang menyoroti kontribusi yang ada serta keterbatasan yang masih ada yang mendorong perlunya penelitian ini.

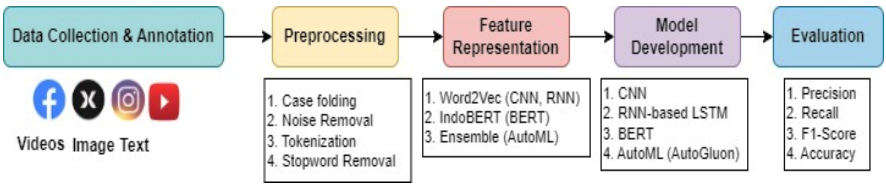
Tabel 9. 1 Ringkasan Studi-Studi Terkait dalam Deteksi Konten Multimodal dan Klasifikasi Teks Antar Model

Studi	**Domain**	**Tipe Data**	**Metode**	**Keunggulan Utama**	**Keterbatasan**
Perdana et al. 2024	Deteksi perjudian Indonesia	Teks (Twitter)	TF-IDF + Random Forest	Akurasi tinggi (~96%) pada iklan perjudian	Generalisasi terbatas, memerlukan desain fitur manual
Thapa et al. 2024	Ujaran kebencian dalam meme	Gambar	OCR + Fitur multimodal	Menangkap teks tersembunyi dalam gambar	Kinerja menurun dengan resolusi rendah atau font bergaya

Xu et al. 2025	Berita palsu video	Frame video	Sampling frame + OCR	Memperluas deteksi ke konten video multimodal	Sensitif terhadap blur gerakan, biaya komputasi tinggi
Bell et al. 2025	Pidato beracun	Audio + teks	Transkripsi ASR + Klasifikasi	Memungkinkan moderasi konten berbicara	Kesalahan transkripsi dari kebisingan, aksen, dan variasi
Adila et al. 2024	ASR Indonesia	Audio (pidato)	Model ASR end-to-end	Membangun sumber daya ASR Indonesia	Kekurangan dataset yang beragam, penanganan kode-switching buruk
Wilie et al. 2020	NLP Indonesia	Teks	IndoBERT	Adaptasi kuat untuk Bahasa Indonesia	Ukuran besar, inferensi lambat
Anitha et al. 2025	Tugas NLP	Teks	AutoGluon (AutoML)	Pemilihan model/parameter otomatis	Masih membutuhkan penyesuaian domain

9.3 Metodologi

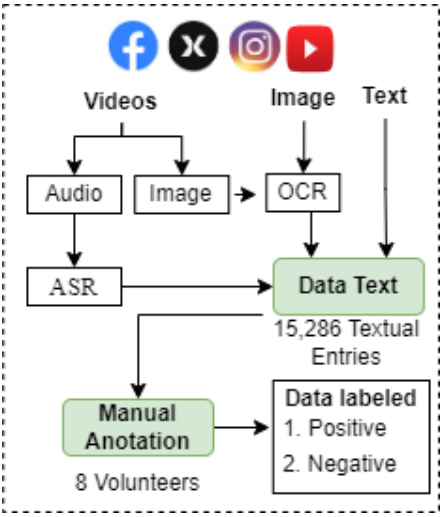
Penelitian ini mengikuti alur penelitian yang terstruktur yang terdiri dari lima tahap utama: (1) pengumpulan dan anotasi data, (2) preprocessing, (3) representasi fitur, (4) pengembangan model, dan (5) evaluasi. Alur kerja keseluruhan ditunjukkan dalam Gambar 9.1. Metodologi yang diusulkan menggabungkan proses ekstraksi teks multimodal untuk menyatukan konten dari berbagai sumber data ke dalam format berbasis teks sebelum dilakukan klasifikasi.



Gambar 9. 1 Alur Kerja Deteksi Iklan Judi

9.3.1 Pengumpulan dan Anotasi Data

Data dikumpulkan dari empat platform media sosial utama: Facebook, X (sebelumnya Twitter), Instagram, dan YouTube, dengan mengikuti alur kerja yang ditunjukkan dalam Gambar 9.2. Proses dimulai dengan pengumpulan data mentah yang mencakup teks pos, gambar, dan video. Pada tahap kedua, konten video diproses dengan mengekstraksi frame kunci pada interval tetap, kemudian diterapkan OCR untuk mengambil elemen teks yang ada. Begitu juga gambar statis, termasuk banner promosi, diproses melalui OCR untuk mengekstrak konten tertulis. Pada tahap ketiga, aliran audio dari video ditranskripsikan menjadi teks menggunakan sistem ASR.



Gambar 9. 2 Alur Proses Ekstraksi Data

Proses ekstraksi multimodal ini memungkinkan inklusi promosi perjudian yang sering muncul dalam format non-teks, yang merupakan taktik penghindaran yang umum digunakan oleh operator perjudian. Tahap akhir mengintegrasikan semua keluaran menjadi satu korpus teks yang terstandarisasi, yang terdiri dari tiga modalitas utama: (1) teks asli yang diperoleh langsung dari keterangan, komentar, dan deskripsi, (2) konten teks yang diekstraksi dari gambar dan frame video melalui OCR, dan (3) transkripsi suara-ke-teks dari audio video. Representasi yang terstandarisasi ini menjadi dasar untuk preprocessing dan pelatihan model selanjutnya.

Setelah pengumpulan dan ekstraksi, korpus yang dihasilkan terdiri dari 15.286 entri teks. Setiap entri dianotasi secara manual oleh delapan relawan terlatih yang memiliki pengalaman atau paparan terhadap aktivitas perjudian online. Dataset ini dilabeli menjadi dua kategori: Positif, yang merujuk pada teks yang secara eksplisit atau implisit mempromosikan atau mendiskusikan aktivitas perjudian online, dan Negatif, yang merujuk pada teks yang tidak terkait dengan perjudian, seperti percakapan umum atau promosi yang tidak relevan.

9.3.2 Preprocessing

Tahap preprocessing bertujuan untuk membersihkan dan menormalkan data teks yang dikumpulkan agar lebih konsisten dan meningkatkan kinerja model. Langkah-langkah preprocessing yang dilakukan meliputi:

1. Case Folding: Mengubah semua karakter menjadi huruf kecil untuk menghindari variasi akibat kapitalisasi huruf.
2. Pembersihan Kebisingan: Menghapus karakter non-alfanumerik seperti emoji, simbol, dan URL.
3. Tokenisasi: Memisahkan kalimat menjadi token individu berdasarkan spasi, dengan penyesuaian untuk memisahkan tanda baca dari kata.
4. Penghapusan Stopword: Menggunakan daftar stopword Sastrawi untuk bahasa Indonesia untuk menghapus kata-kata yang sering muncul tetapi tidak memberikan informasi penting.

9.3.3 Representasi Fitur

Dua strategi embedding digunakan dalam penelitian ini: Word2Vec dan IndoBERT.

- CNN dan RNN (LSTM): Embedding Word2Vec (vektor berdimensi 200 yang telah dipra-latih pada korpora besar bahasa Indonesia) mempertahankan hubungan semantik, menghasilkan urutan kalimat yang dimasukkan ke dalam arsitektur neural.

- Model berbasis BERT: IndoBERT-base-p1 menghasilkan embedding kontekstual melalui tokenisasi WordPiece, dengan urutan yang dipadatkan atau dipotong hingga panjang 128. Berbeda dengan embedding statis, IndoBERT menghasilkan vektor dinamis yang sensitif terhadap konteks.
- AutoML (AutoGluon): Menangani representasi fitur secara otomatis dengan preprocessing internal (tokenisasi, BoW, TF-IDF, n-grams) dan embedding pra-latihan dalam ensemble.

9.3.4 Pengembangan Model

Pada studi kasus ini mengimplementasikan dan membandingkan empat pendekatan model untuk mengklasifikasikan iklan perjudian online pada konten media sosial berbahasa Indonesia. Dua pendekatan menggunakan model deep learning yang dirancang secara manual (CNN dan RNN dengan LSTM), satu model berbasis transformer (IndoBERT), dan satu pendekatan AutoML (AutoGluon). Semua model dilatih dan dievaluasi menggunakan dataset yang sama dan metrik kinerja yang seragam untuk memastikan perbandingan yang adil. Pada Tabel 9.2 menunjukkan arsitektur dan pengaturan parameter yang digunakan dalam kasus ini.

Tabel 9. 2 Arsitektur dan Parameter Pengklasifikasi

Model	Arsitektur dan Parameter
CNN-Based Classifier	1. Layer embedding: Word2Vec pra-latihan, vektor 200-dim 2. Conv1D: 128 filter, ukuran kernel = 5, aktivasi = ReLU 3. Max-Pooling layer 4. Dense layer: fully connected, dropout = 0.5 5. Output: sigmoid layer 6. Loss: Binary Cross-Entropy 7. Optimizer: Adam, LR = 0.001 8. Pelatihan: 10 epoch, ukuran batch = 64, early stopping
RNN-Based Classifier (LSTM)	1. Layer embedding: Word2Vec pra-latihan, vektor 200-dim 2. LSTM layer: 128 unit tersembunyi 3. Dense layer: 64 unit, aktivasi = ReLU 4. Dropout = 0.5 5. Output: sigmoid layer 6. Loss: Binary Cross-Entropy 7. Optimizer: Adam, LR = 0.001 8. Pelatihan: 10 epoch, ukuran batch = 64, early stopping
Transformer-Based Classifier (IndoBERT)	1. Tokenizer: WordPiece, panjang maksimal = 128 2. Encoder: indobert-base-p1 3. Dense layer: dropout = 0.3 (pooled output) 4. Output: sigmoid layer 5. Loss: Binary Cross-Entropy

	6. Optimizer: AdamW, LR = 2e-5 7. Pelatihan: 4 epoch, ukuran batch = 32
AutoML (AutoGluon)	1. Framework: AutoGluon 2. Preprocessing otomatis: tokenisasi, ekstraksi fitur, penanganan nilai yang hilang 3. Model kandidat: GBM, Random Forest, Bagging Ensembles, Feed-forward NN 4. Pembuatan ensemble bertumpuk 5. Waktu pelatihan: maksimum 300 detik per percobaan

9.3.5 Evaluasi

Untuk mengevaluasi efektivitas model dalam mendeteksi iklan perjudian online, penelitian ini menggunakan metrik klasifikasi standar yang umum digunakan dalam tugas klasifikasi teks biner: akurasi, presisi, recall, dan F1-score. Metrik ini memberikan penilaian kinerja yang komprehensif dari berbagai perspektif, menangkap baik kebenaran maupun ketahanan dalam klasifikasi. Selain evaluasi numerik, analisis kesalahan dilakukan dengan meninjau sebagian data yang salah diklasifikasikan, untuk mengidentifikasi pola linguistik atau petunjuk kontekstual yang berkontribusi pada kesalahan prediksi.

9.4 Hasil dan Diskusi

9.4.1 Pengaturan Eksperimen dan Tuning Hyperparameter

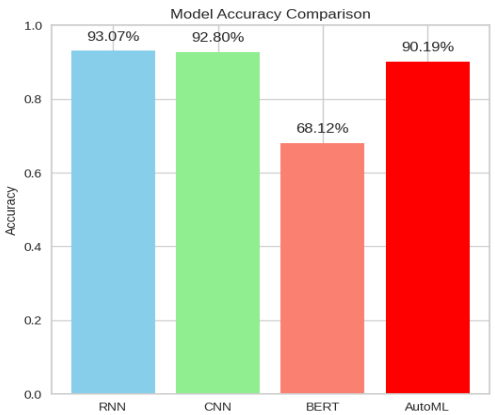
Pada studi kasus ini, serangkaian eksperimen terkontrol dilakukan untuk mengevaluasi kinerja kerangka kerja multimodal-to-text yang diusulkan dalam mendeteksi iklan perjudian online di media sosial Indonesia. Tujuan utama adalah untuk menentukan konfigurasi optimal dari model deep learning dan berbasis AutoML, memastikan kinerja klasifikasi yang kuat dalam menghadapi tantangan linguistik dan kontekstual. Dua skenario eksperimen dirancang:

1. Evaluasi Arsitektur Model: Membandingkan kinerja model CNN, LSTM, BERT, dan AutoML di bawah kondisi preprocessing dan dataset yang terstandarisasi.
2. Optimasi Hyperparameter: Menemukan set hyperparameter terbaik untuk setiap model agar mencapai akurasi dan F1-score maksimal.

Sebuah pipeline preprocessing yang terstandarisasi diterapkan pada semua model, termasuk tokenisasi, case-folding, penghapusan stopword, serta penanganan emoji, slang, dan kode-switching. Hal ini memastikan bahwa perbedaan kinerja yang tercatat murni disebabkan oleh kemampuan model, bukan variasi dalam preprocessing. Hyperparameter yang diuji meliputi batch size, learning rate, dan jumlah epoch. Pengaturan optimal untuk CNN dan LSTM ditemukan dengan batch size 64, learning rate 0.001, dan 10 epoch, sementara BERT mencapai hasil terbaik dengan batch size 32, learning rate 5e-5, dan 5 epoch. Untuk AutoML, konfigurasi terbaik dicapai dengan pengaturan waktu pelatihan maksimum 300 detik per percobaan.

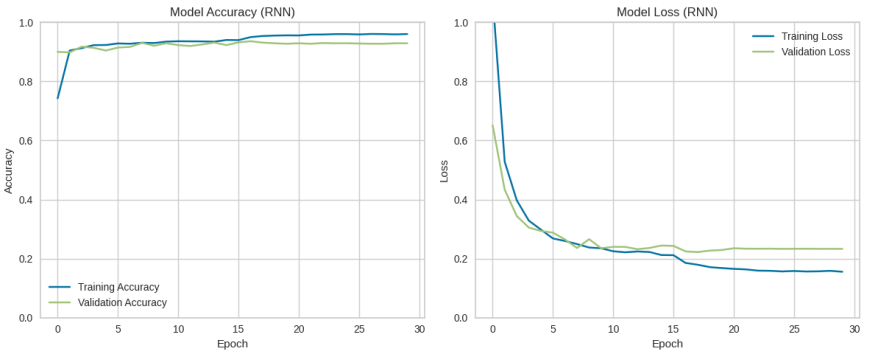
9.4.2 Perbandingan Akurasi Model dan Analisis Kurva Pembelajaran

Untuk mengevaluasi kinerja relatif model yang diusulkan, dilakukan analisis perbandingan pada akurasi klasifikasi yang dicapai oleh RNN, CNN, BERT, dan AutoML di bawah konfigurasi hyperparameter optimal mereka. Hasilnya, model RNN mencapai akurasi tertinggi yaitu 93,07%, diikuti oleh CNN dengan 92,80% dan AutoML dengan 90,19%, sementara BERT memperoleh akurasi yang lebih rendah, yaitu 68,12%. Hasilnya kinerja dapat dilihat pada Gambar 9.3.



Gambar 9. 3 Hasil Kinerja Model

Kinerja superior dari RNN dapat dikaitkan dengan kemampuannya dalam menangkap ketergantungan sekuensial jangka panjang dalam teks, yang sangat berguna dalam mengidentifikasi pola kontekstual dalam iklan perjudian. CNN juga menunjukkan kinerja kompetitif, efektif dalam mengekstrak pola n-gram lokal dari teks, meskipun sedikit kurang efektif dibandingkan RNN dalam memodelkan ketergantungan kontekstual yang lebih panjang. AutoML menunjukkan kinerja yang stabil tanpa penyetelan arsitektur manual, memanfaatkan seleksi fitur otomatis dan strategi ensemble. Sementara itu, BERT menunjukkan akurasi rendah, mengindikasikan bahwa adaptasi domain khusus diperlukan untuk memaksimalkan potensi embedding kontekstual yang telah dilatih sebelumnya. Kemudian untuk melihat grafik kinerja pelatihan dan validasi dapat dilihat pada Gambar 9.4



Gambar 9. 4 Grafik Kinerja Akurasi dan Validasi Model

9.4.3 Analisis Kesalahan

Analisis kesalahan dilakukan untuk memahami kekuatan dan kelemahan setiap model dalam mendeteksi iklan perjudian online di media sosial Indonesia. Berdasarkan metrik kinerja per kelas, model RNN mencapai hasil yang paling seimbang dengan presisi, recall, dan F1-score lebih dari 93% untuk kedua kelas, Positif dan Negatif. Model CNN juga memberikan hasil yang kuat namun dengan sedikit ketidakstabilan dalam recall untuk kelas Negatif. AutoML menunjukkan kinerja yang relatif seimbang di seluruh kelas,

sedangkan BERT mengalami penurunan signifikan dengan F1-score sekitar 68% untuk kedua kelas.

Selain analisis numerik, dilakukan pula analisis confusion matrix yang menunjukkan bahwa model RNN memiliki jumlah false positives dan false negatives yang seimbang, mengindikasikan kemampuan prediktif yang baik. CNN menghasilkan false positives lebih banyak namun lebih sensitif terhadap konten perjudian. AutoML juga menunjukkan hasil kompetitif, meskipun false negatives-nya lebih tinggi, yang menunjukkan kecenderungan untuk melewatkan kasus positif. BERT, di sisi lain, menunjukkan kinerja terlemah dengan jumlah false positives dan false negatives yang jauh lebih tinggi, yang mencerminkan kesulitan dalam membedakan konten perjudian dari non-perjudian.

9.4.4 Diskusi Perbandingan

Evaluasi perbandingan antara RNN, CNN, BERT, dan AutoML menyoroti kekuatan dan keterbatasan masing-masing pendekatan dalam mendeteksi iklan perjudian di media sosial Indonesia. RNN berhasil mengatasi tantangan bahasa informal dan kode-switching dengan baik, berkat kemampuannya dalam menangkap ketergantungan sekuensial. CNN efektif dalam mendeteksi konten perjudian yang kaya kata kunci, tetapi cenderung menghasilkan false positives dalam konteks ambigu. BERT, meskipun kuat dalam pemahaman konteks, menunjukkan kinerja buruk karena kurangnya adaptasi terhadap bahasa Indonesia yang penuh dengan slang dan noise. AutoML, meskipun memberikan hasil yang cukup seimbang, terbatas dalam menangani nuansa linguistik yang spesifik pada domain perjudian.

Dibandingkan dengan studi sebelumnya, kerangka kerja multimodal-to-text yang diusulkan menunjukkan peningkatan yang jelas, dengan menggabungkan sinyal multimodal (teks, gambar, dan audio), yang memungkinkan generalisasi lebih baik di berbagai platform. Temuan ini menunjukkan bahwa meskipun BERT memiliki keunggulan dalam model berbasis Transformer, pendekatan seperti RNN dan CNN lebih efektif dalam konteks data yang penuh noise dan

bahasa informal, mengingat kelebihan mereka dalam menangani dependensi sekuensial dan pola n-gram lokal.

9.5 Kesimpulan

Studi kasus ini berhasil mengembangkan dan mengevaluasi kerangka kerja multimodal-to-text untuk mendeteksi iklan perjudian online di media sosial Indonesia dengan menggabungkan berbagai sumber data, termasuk teks asli, gambar, dan audio. Kerangka kerja ini menawarkan pendekatan yang lebih komprehensif dibandingkan dengan sistem deteksi tradisional yang hanya mengandalkan teks. Berdasarkan hasil eksperimen, model RNN menunjukkan kinerja terbaik dengan akurasi mencapai 93,07%, diikuti oleh CNN dan AutoML. Sementara itu, BERT mengalami penurunan kinerja yang signifikan, mengindikasikan perlunya adaptasi domain yang lebih mendalam untuk menangani bahasa Indonesia yang penuh dengan slang dan kebisingan.

Secara keseluruhan, model RNN menonjol berkat kemampuannya dalam menangkap ketergantungan sekuensial dan pola kontekstual dalam teks yang lebih panjang, sementara CNN unggul dalam mendeteksi iklan perjudian yang kaya kata kunci. Namun, meskipun AutoML memberikan hasil yang seimbang, model ini memiliki keterbatasan dalam menangani nuansa linguistik spesifik yang berkaitan dengan perjudian. BERT, meskipun memiliki potensi dalam pemahaman konteks, terbukti tidak optimal dalam mendeteksi iklan perjudian di media sosial Indonesia tanpa penyesuaian domain yang lebih kuat.

Kerangka kerja multimodal menunjukkan bahwa integrasi teks, gambar, dan audio dapat memperbaiki deteksi konten perjudian yang tersembunyi, memberikan solusi yang lebih kuat dan dapat diperluas untuk moderasi konten di platform media sosial. Namun, tantangan seperti kebisingan dalam OCR/ASR dan ambiguitas kontekstual tetap menjadi masalah yang perlu ditangani dalam penelitian selanjutnya. Pengembangan lebih lanjut di bidang ini, seperti perluasan kosakata domain dan peningkatan model berbasis transformer untuk menangani teks berbahasa campuran.

BAB X
PENUTUP

Perkembangan media sosial yang sangat cepat telah membawa manfaat besar sekaligus tantangan serius bagi masyarakat. Hoaks, ujaran kebencian, pornografi, promosi judi online, penipuan digital, hingga berbagai bentuk konten manipulatif lainnya menjadikan ruang digital rentan terhadap penyalahgunaan. Dalam konteks inilah, moderasi konten berbasis kecerdasan buatan (AI) hadir sebagai salah satu pilar penting untuk menjaga ekosistem media sosial tetap aman, sehat, dan produktif. Buku ini disusun untuk memberikan gambaran menyeluruh mengenai bagaimana AI dapat dimanfaatkan secara strategis dan beretika dalam proses moderasi tersebut.

Secara garis besar, Bagian I buku ini telah mengantarkan pembaca memahami lanskap masalah: bagaimana media sosial berkembang, jenis-jenis konten berbahaya yang muncul, serta urgensi moderasi konten bagi keamanan digital dan kohesi sosial. Pembaca diajak melihat bahwa moderasi bukan sekadar aktivitas teknis, tetapi berkaitan erat dengan perlindungan generasi muda, stabilitas masyarakat, hingga kepercayaan publik terhadap platform digital.

Bagian II kemudian memaparkan fondasi teknis kecerdasan buatan yang relevan dengan moderasi konten. Di sini dijelaskan perbedaan pendekatan machine learning dan deep learning, peran Natural Language Processing (NLP), Computer Vision (CV), dan Speech Recognition, serta berbagai algoritma populer yang kerap digunakan dalam sistem moderasi, mulai dari metode klasik (Naive Bayes, SVM, Random Forest) hingga model deep learning modern seperti CNN, RNN, LSTM, dan Transformer. Pembahasan dilengkapi dengan konsep AutoML dan transfer learning, termasuk pemanfaatan model bahasa seperti BERT dan IndoBERT yang sangat penting untuk menangkap konteks lokal, khususnya dalam bahasa Indonesia.

Bagian III menempatkan seluruh konsep tersebut dalam konteks aplikasi nyata. Moderasi konten teks melalui deteksi ujaran kebencian, hoaks, dan promosi ilegal menunjukkan bagaimana NLP

dan model bahasa dapat bekerja untuk memfilter percakapan digital yang sangat dinamis. Moderasi konten visual (gambar dan video) menggambarkan pemanfaatan Computer Vision dan OCR untuk mengenali konten eksplisit maupun teks terselubung dalam media visual. Sementara itu, moderasi konten audio dan pendekatan multimodal memperlihatkan bagaimana ASR, NLP, dan CV dapat digabungkan untuk menganalisis teks, gambar, dan suara secara terpadu. Pendekatan multimodal ini menegaskan bahwa pelanggaran kebijakan platform sering kali hadir secara halus dan tersebar di berbagai bentuk data sekaligus.

Dari seluruh pembahasan, terdapat beberapa pesan kunci yang dapat digarisbawahi. Pertama, moderasi konten digital adalah persoalan multidimensi: teknis, sosial, hukum, dan etika. AI dapat membantu mengatasi skala dan kompleksitas data, tetapi tidak dapat berdiri sendiri. Keterlibatan manusia tetap dibutuhkan dalam merancang kebijakan, mengawasi keputusan model, serta menangani kasus-kasus di wilayah abu-abu yang memerlukan sensitivitas nilai dan konteks budaya. Kedua, kualitas data adalah faktor penentu keberhasilan sistem moderasi. Prinsip "garbage in, garbage out" menegaskan bahwa tanpa data yang baik, representatif, dan teranotasi dengan cermat, model AI berisiko bias, tidak adil, atau justru memperkuat stereotipe. Ketiga, transparansi dan akuntabilitas harus menjadi bagian integral dari desain sistem, agar pengguna memahami mengapa suatu konten ditandai atau dihapus, serta memiliki mekanisme banding yang jelas.

Ke depan, terdapat sejumlah tantangan dan peluang yang perlu mendapat perhatian. Bahasa dan budaya digital terus berkembang; istilah baru, bahasa gaul, dan strategi penyamaran konten berbahaya akan selalu muncul. Hal ini menuntut sistem moderasi yang adaptif, mampu belajar secara berkelanjutan, dan sensitif terhadap konteks lokal. Di saat yang sama, tuntutan terhadap perlindungan privasi, keamanan data, dan keadilan algoritmik (algorithmic fairness) juga semakin menguat. Penelitian-penelitian baru diperlukan untuk memastikan bahwa model yang digunakan tidak mendiskriminasi kelompok tertentu, tidak memperkuat bias

yang ada, dan tetap menghormati kebebasan berekspresi dalam batas-batas yang bertanggung jawab.

Berdasarkan itu, beberapa arah pengembangan dan rekomendasi dapat dirangkum sebagai berikut:

1. Bagi akademisi dan peneliti, perlu terus dikembangkan:
    - Dataset lokal yang mencerminkan kekayaan bahasa dan budaya Indonesia, termasuk slang, kode-kode promosi ilegal, serta variasi dialek.
    - Model-model AI yang lebih transparan dan dapat dijelaskan (explainable AI) agar keputusan moderasi dapat diaudit dan dipertanggungjawabkan.
    - Kajian lintas disiplin yang menggabungkan perspektif ilmu komputer, hukum, komunikasi, psikologi, dan studi kebijakan publik.
2. Bagi praktisi industri dan pengelola platform, penting untuk:
    - Membangun arsitektur sistem moderasi yang end-to-end, mulai dari deteksi otomatis, verifikasi manual, hingga pelaporan dan analitik kebijakan.
    - Melakukan evaluasi berkala terhadap kinerja model, termasuk memantau metrik seperti false positive dan false negative yang berdampak langsung pada pengalaman pengguna.
    - Berkolaborasi dengan komunitas lokal, peneliti, dan regulator untuk memastikan bahwa kebijakan serta implementasi moderasi selaras dengan nilai-nilai masyarakat dan peraturan yang berlaku.
3. Bagi pemerintah dan pembuat kebijakan, dibutuhkan:
    - Kerangka regulasi yang adaptif terhadap perkembangan teknologi, yang tidak menghambat inovasi tetapi tetap melindungi warga dari dampak negatif konten berbahaya.
    - Panduan etika dan standar nasional terkait penggunaan AI dalam moderasi konten, termasuk aspek transparansi, akuntabilitas, dan perlindungan data pribadi.

- Fasilitasi ekosistem kolaboratif antara kampus, industri, dan komunitas untuk mengembangkan solusi moderasi konten yang berkelanjutan.

4. Bagi masyarakat luas dan pengguna media sosial, peran aktif sangat dibutuhkan melalui:
    - Peningkatan literasi digital dan kemampuan berpikir kritis, sehingga tidak mudah terprovokasi oleh informasi yang menyesatkan.
    - Pemanfaatan fitur pelaporan (report) dan mekanisme umpan balik di platform untuk membantu deteksi konten berbahaya.
    - Budaya berbagi informasi yang bertanggung jawab, dengan memeriksa kebenaran sebelum menyebarkan suatu konten.

Pada akhirnya, keberhasilan moderasi konten digital tidak hanya ditentukan oleh kecanggihan algoritma, tetapi oleh sejauh mana seluruh pemangku kepentingan mampu bekerja bersama membangun ruang digital yang aman, inklusif, dan manusiawi. Kecerdasan buatan hanyalah alat; tujuan utamanya tetaplah perlindungan martabat manusia dan kualitas kehidupan bermasyarakat di era informasi.

Semoga buku ini dapat menjadi pijakan awal yang kokoh untuk melangkah ke arah yang lebih lanjut menuju pemahaman dan pengembangan sistem moderasi konten digital berbasis kecerdasan buatan yang lebih matang, adil, dan berkelanjutan. Harapannya, buku ini tidak hanya menjadi bahan bacaan, tetapi juga pemicu lahirnya penelitian baru, inovasi teknologi, serta kebijakan publik yang lebih baik dalam menghadapi tantangan konten digital di masa depan.

DAFTAR PUSTAKA

Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity-A Survey. *ACM Computing Surveys*, *54*(2), 1–35. https://doi.org/10.1145/3440755

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805

Gongane, V. U., Munot, M. V, & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, *12*(1), 129.

Irawan, Yusufianto, Agustina, D. (2020). Laporan Survei Internet APJII 2019 – 2020. *Asosiasi Penyelenggara Jasa Internet Indonesia*, *2020*, 1–146. https://apjii.or.id/survei

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770.

Lopez, M. M., & Kalita, J. (2017). *Deep Learning applied to NLP*. http://arxiv.org/abs/1703.03091

Murugesan, S., & Kaliyamurthie, K. P. (2023). A Machine Learning Framework for Automatic Fake News Detection in Indian Tamil News Channels. *Ingénierie Des Systèmes d Information*, *28*(1), 205–209. https://doi.org/10.18280/isi.280123

Muzakir, A. (2024). *Penerapan Konsep Machine Learning & Deep Learning*. Asosiasi Dosen Sistem Informasi Indonesia.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: a modern approach*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

BIODATA PENULIS

Dr. Ari Muzakir, S.Kom., M.Cs, merupakan dosen di Program Studi Teknik Informatika, Fakultas Sains Teknologi, Universitas Bina Darma. Gelar Magister Ilmu Komputer diperoleh dari Universitas Gadjah Mada tahun 2012 dan gelar Doktor Sistem Informasi dari Universitas Diponegero tahun 2024. Spesialisasi dalam bidang Pemrosesan Bahasa Alami (NLP), yang berkontribusi pada kemajuan dalam memahami dan memproses bahasa manusia melalui metode komputasi. Email: arimuzakir@binadarma.ac.id.

Dr. Usman Ependi, S.Kom, M.Kom, saat ini sebagai Ketua Program Magister Teknik Informatika, Universitas Bina Darma. Ia meraih gelar sarjana dalam Ilmu Komputer pada tahun 2009 dan gelar magister Komputer tahun 2011 di Universitas Bina Darma. Kemudian meraih gelar Doktor sistem informasi di Universitas Diponegoro tahun 2024. Fokus pada riset di bidang rekayasa perangkat lunak dan kecerdasan buatan, dengan visi untuk mengembangkan konsep inovatif untuk smart city. email: u.ependi@binadarma.ac.id.

Suyanto, S.Kom., M.M., M.Kom merupakan dosen Program Studi Sistem Informasi (S1), Fakultas Sains dan Teknologi, Universitas Bina Darma. Ia menyelesaikan pendidikan Sarjana Komputer tahun 2000, gelar Magister Manajemen pada tahun 2006, dan gelar Magister Komputer pada tahun 2011 di Universitas Bina Darma. Bidang minat penelitian berfokus pada rekayasa perangkat lunak sistem informasi. Email: suyanto@binadarma.ac.id.

Buku ini membahas peran kecerdasan buatan (AI) dalam moderasi konten digital di media sosial, sebuah topik yang semakin relevan di era informasi saat ini. Buku ini mengulas berbagai jenis konten berbahaya seperti hoaks, ujaran kebencian, pornografi, dan penipuan digital, serta urgensi moderasi untuk memastikan ruang digital tetap aman. Dengan memanfaatkan teknologi AI, buku ini mengeksplorasi bagaimana sistem dapat dilatih untuk mendeteksi dan menanggulangi konten berbahaya secara otomatis dengan tingkat akurasi tinggi, yang tentunya sangat penting untuk menjaga keberlanjutan dan kesehatan ekosistem digital.

Melalui penjelasan yang komprehensif, buku ini menyajikan pendekatan teknis dalam moderasi konten berbasis AI, mulai dari dasar-dasar machine learning, deep learning, hingga penerapan algoritma canggih seperti Natural Language Processing (NLP), Computer Vision (CV), dan Speech Recognition. Pembaca akan dipandu untuk memahami cara-cara model AI digunakan untuk mendeteksi ujaran kebencian, hoaks, konten visual eksplisit, dan bahkan konten audio yang melanggar aturan platform media sosial.

Di sisi lain, buku ini juga membahas berbagai tantangan yang muncul dalam penerapan AI untuk moderasi konten, termasuk masalah bias dalam algoritma, transparansi, dan akuntabilitas sistem. Secara keseluruhan, buku ini bukan hanya memberikan wawasan tentang teknologi moderasi konten berbasis AI, tetapi juga mengajak pembaca untuk berpikir kritis tentang etika, privasi, dan dampak sosial dari teknologi tersebut dalam menjaga kebebasan berekspresi di dunia digital.

Diterbitkan oleh

ASOSIASI DOKTOR Sistem Informasi INDONESIA

ISBN 978-634-04-6360-6

9 786340 463606