ANALISIS SENTIMEN

Konsep dan Implementasi

Suyahman, S.Kom., M.Kom.
Yogiek Indra Kurniawan, S.T., M.T.
Muhammad Adi Pratama, S.S., M.A.
Deny Prasetyo, S.Pd., M.Kom.
Nurul Badriah, S.Kom., M.Kom.
Ardy Wicaksono, S.Kom., M.Kom.
Muhammad Anwar Fauzi, S.Kom., M.Kom.
Satria Pradana Rizki Yulianto, M.Kom.
Ayun Hapsari, S.Kom.
Egi Dio Bagus Sudewo, S.Kom., M.Kom.
Muhammad Oktoda Noorrohman S.Kom., M.Tr.T.

SAH PUBLISHER
PT SOLUSI ADMINISTRASI HUKUM

Analisis Sentimen

Konsep dan Implementasi

UU No. 28 Tahun 2014 tentang Hak Cipta

Fungsi dan Sifat Hak Cipta (Pasal 4)

Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Perlindungan (Pasal 26)

Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

1. Penggunaan kutipan singkat ciptaan dan/atau produk hak terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk penyediaan informasi aktual.
2. Penggandaan ciptaan dan/atau produk hak terkait hanya untuk kepentingan penelitian ilmu pengetahuan.
3. Penggandaan ciptaan dan/atau produk hak terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan fonogram yang telah dilakukan pengumuman sebagai bahan ajar.
4. Penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu ciptaan dan/atau produk hak terkait digunakan tanpa izin pelaku pertunjukan, produser fonogram, atau lembaga penyiaran.

Sanksi Pelanggaran (Pasal 113)

1. Setiap orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk penggunaan secara komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp100.000.000,00 (seratus juta rupiah).
2. Setiap orang yang dengan tanpa hak dan/atau tanpa izin pencipta atau pemegang hak cipta melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk penggunaan secara komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

Analisis Sentimen

Konsep dan Implementasi

Suyahman, S.Kom., M.Kom.
Yogiek Indra Kurniawan, S.T., M.T.
Muhammad Adi Pratama, S.S., M.A.
Deny Prasetyo, S.Pd., M.Kom.
Nurul Badriah, S.Kom., M.Kom.
Ardy Wicaksono, S.Kom., M.Kom.
Muhammad Anwar Fauzi, S.Kom., M.Kom.
Satria Pradana Rizki Yulianto, M.Kom.
Ayun Hapsari, S.Kom.
Egi Dio Bagus Sudewo, S.Kom., M.Kom.
Muhammad Oktoda Noorrohman, S.Kom., M.Tr.T.

PT Solusi Administrasi Hukum

Analisis Sentimen
Konsep dan Implementasi

Nama Penulis
Suyahman, S.Kom., M.Kom.
Yogiek Indra Kurniawan, S.T., M.T.
Muhammad Adi Pratama, S.S., M.A.
Deny Prasetyo, S.Pd., M.Kom.
Nurul Badriah, S.Kom., M.Kom.
Ardy Wicaksono, S.Kom., M.Kom.
Muhammad Anwar Fauzi, S.Kom., M.Kom.
Satria Pradana Rizki Yulianto, M.Kom.
Ayun Hapsari, S.Kom.
Egi Dio Bagus Sudewo, S.Kom., M.Kom.
Muhammad Oktoda Noorrohman S.Kom., M.Tr.T.

Desain Cover :
Tim Sah Publisher

Ilustrasi :
Gemini

Tata Letak :
Tim Sah Publisher

Ukuran :
146 halaman, Uk: 17.5x25 cm

E-ISBN :
XXX-X-XXXXX-XXX-X

Cetakan Pertama :
Februari 2026

PENERBIT PT SOLUSI ADMINISTRASI HUKUM

Jl. Pedan Karangdowo, Munggung, Karangdowo, Klaten
Telp: +628562160034
publisher.sah.co.id
E-mail: publisher@sah.co.id

KATA PENGANTAR

Perkembangan teknologi digital telah mengubah cara masyarakat menyampaikan opini. Setiap hari, jutaan teks diproduksi melalui media sosial, ulasan produk, forum diskusi, dan survei daring. Data tersebut mengandung informasi berharga tentang persepsi, kepuasan, dan respons publik terhadap berbagai isu. Analisis sentimen hadir sebagai pendekatan sistematis untuk memahami pola opini dalam skala besar.

Buku ini disusun untuk memberikan pemahaman yang utuh mengenai analisis sentimen, baik dari sisi teoretis maupun teknis. Pembahasan dimulai dari konsep dasar linguistik dan representasi teks, kemudian berkembang ke metode machine learning dan deep learning. Setiap bab dirancang agar pembaca memahami alur logis dari pengolahan data hingga implementasi sistem. Pendekatan ini membantu pembaca melihat keterkaitan antara teori, model komputasi, dan penerapan praktis.

Materi dalam buku ini menekankan keseimbangan antara fondasi konseptual dan praktik implementasi. Penjelasan tentang evaluasi model, validasi, dan deployment disertakan untuk memastikan bahwa pembaca tidak hanya memahami algoritma, tetapi juga mampu membangun sistem yang berfungsi dengan baik. Pembahasan disusun secara sistematis agar mudah diikuti oleh mahasiswa, peneliti, maupun praktisi.

Buku ini diharapkan dapat menjadi referensi yang bermanfaat dalam pengembangan penelitian dan aplikasi analisis sentimen di berbagai bidang. Setiap bagian dirancang untuk mendorong pembaca berpikir kritis dan memahami batasan metode yang digunakan. Kritik dan masukan dari pembaca akan sangat berharga untuk penyempurnaan edisi berikutnya.

Semoga buku ini dapat memberikan kontribusi nyata dalam pengembangan kajian analisis sentimen dan penerapannya di lingkungan akademik maupun industri.

Klaten, 23 Februari 2026

Penulis

DAFTAR ISI

BAB 1. PENGANTAR ANALISIS SENTIMEN

Suyahman, S.Kom., M.Kom.

Perkembangan teknologi digital mendorong manusia untuk mengekspresikan opini secara terbuka melalui berbagai platform daring. Ulasan produk, komentar berita, dan tanggapan kebijakan publik muncul setiap hari dalam jumlah besar. Kondisi ini menghadirkan peluang sekaligus tantangan bagi peneliti dan praktisi. Analisis sentimen hadir untuk menjawab kebutuhan memahami opini tersebut secara sistematis dan terukur.

A. LATAR BELAKANG DAN URGENSI ANALISIS SENTIMEN

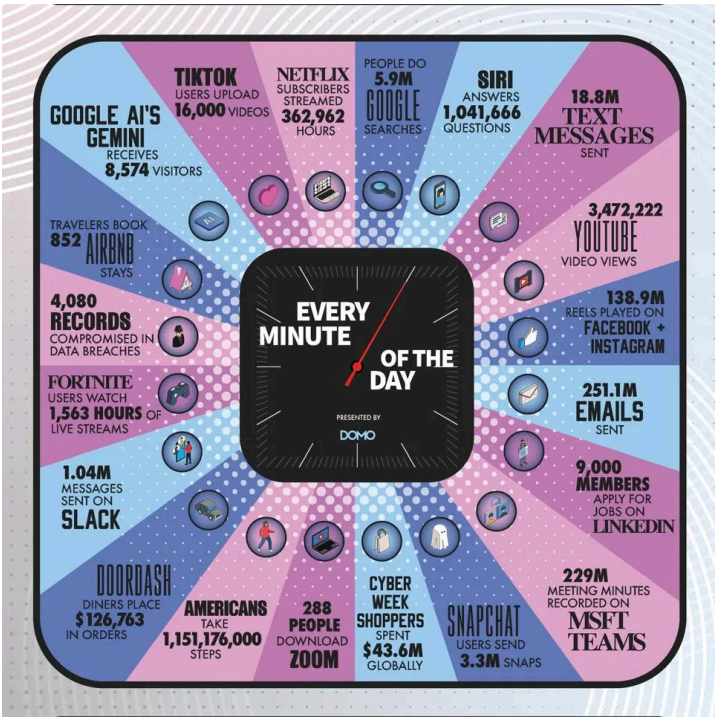
Transformasi digital mengubah cara masyarakat memproduksi dan menyebarkan informasi. Sejak pertengahan 2000-an, pertumbuhan platform seperti Twitter dan Facebook mempercepat arus komunikasi publik. Pengguna tidak lagi berperan sebagai konsumen informasi semata, tetapi juga sebagai produsen opini. Setiap hari jutaan teks pendek dan panjang dipublikasikan dalam bentuk komentar, ulasan, dan diskusi daring.

Ledakan data teks ini menciptakan sumber informasi yang sangat kaya. Perusahaan memanfaatkan ulasan pelanggan untuk menilai kualitas produk. Pemerintah memantau respons masyarakat terhadap kebijakan publik. Perguruan tinggi menganalisis umpan balik mahasiswa untuk meningkatkan layanan akademik. Namun volume data yang sangat besar membuat analisis manual menjadi tidak efektif. Peneliti tidak mungkin membaca ribuan atau jutaan komentar secara satu per satu dalam waktu singkat.

Media sosial memperkuat dinamika produksi opini publik. Platform seperti Instagram dan YouTube menyediakan ruang interaksi yang cepat dan terbuka. Opini dapat menyebar luas hanya dalam hitungan menit. Sebuah isu lokal dapat berubah menjadi diskusi nasional dalam satu hari. Situasi ini menuntut metode analisis yang mampu menangkap kecenderungan sikap publik secara cepat dan akurat.

Tantangan utama terletak pada skala dan kompleksitas bahasa. Bahasa manusia bersifat ambigu dan kontekstual. Satu kata dapat memiliki makna berbeda dalam situasi berbeda. Sarkasme dan ironi sering muncul dalam

komentar daring. Pengguna juga menggunakan singkatan, slang, dan campuran bahasa. Kompleksitas ini menyulitkan proses interpretasi otomatis jika tidak dirancang dengan pendekatan yang tepat.



Ilustrasi banyaknya Data yang Dihasilkan Internet Dalam Satu Menit (https://bondhighplus.com/blog/what-happens-in-an-internet-minute/)

Kondisi tersebut mendorong kebutuhan akan otomatisasi analisis opini. Sistem komputasional mampu memproses ribuan dokumen dalam waktu singkat. Algoritma pembelajaran mesin dapat mengenali pola sentimen dari data pelatihan. Model deep learning bahkan mampu menangkap konteks kalimat yang lebih kompleks. Otomatisasi tidak hanya mempercepat proses, tetapi juga meningkatkan konsistensi hasil analisis.

Relevansi analisis sentimen semakin kuat dalam dunia penelitian dan industri. Dalam penelitian sosial, analisis ini membantu memahami persepsi publik terhadap isu tertentu. Dalam bidang bisnis, perusahaan menggunakan hasil analisis untuk menyusun strategi pemasaran. Di sektor keuangan, investor memanfaatkan sentimen pasar untuk memprediksi pergerakan saham. Analisis sentimen tidak lagi menjadi sekadar alat teknis, tetapi telah menjadi bagian dari pengambilan keputusan berbasis data.

Dengan latar belakang tersebut, kajian tentang analisis sentimen menjadi penting untuk dipahami secara konseptual dan metodologis. Bab ini menjadi dasar untuk memahami bagaimana opini digital dapat diolah menjadi informasi yang bermakna dan dapat dipertanggungjawabkan secara ilmiah.

B. DEFINISI DAN RUANG LINGKUP ANALISIS SENTIMEN

Analisis sentimen merupakan teknik komputasional yang bertujuan mengidentifikasi sikap atau opini dalam teks. Teknik ini memproses data bahasa untuk menentukan kecenderungan positif, negatif, atau netral. Penelitian ini memanfaatkan metode pemrosesan bahasa alami dan pembelajaran mesin untuk mengekstraksi makna dari teks tidak terstruktur. Fokus utama terletak pada interpretasi sikap penulis terhadap suatu objek, peristiwa, atau isu tertentu.

Istilah analisis sentimen sering digunakan bersama dengan opinion mining. Keduanya memiliki hubungan erat, tetapi tidak sepenuhnya sama. Opinion mining menekankan proses ekstraksi opini secara umum, termasuk identifikasi siapa yang berpendapat dan terhadap apa pendapat tersebut diarahkan. Analisis sentimen lebih spesifik pada pengukuran orientasi atau kecenderungan emosional dari opini tersebut. Perbedaan keduanya dapat diringkas dalam tabel berikut.

Aspek	Opinion Mining	Analisis Sentimen
Fokus utama	Ekstraksi opini secara menyeluruh	Identifikasi kecenderungan sentimen
Elemen yang dianalisis	Penulis, target, dan isi opini	Polarity dan nilai sentimen
Ruang lingkup	Lebih luas dan konseptual	Lebih terfokus pada klasifikasi
Tujuan akhir	Memetakan struktur opini	Menentukan nilai positif, negatif, atau netral

Dalam praktiknya, klasifikasi sentimen umumnya dibagi menjadi tiga kategori dasar, yaitu positif, negatif, dan netral. Kategori positif menunjukkan dukungan atau kepuasan terhadap suatu objek. Kategori negatif menunjukkan penolakan atau ketidakpuasan. Kategori netral menggambarkan pernyataan yang bersifat informatif tanpa muatan emosional yang jelas. Pembagian ini menjadi fondasi dalam banyak penelitian awal di bidang ini.

Selain klasifikasi dasar, kajian ini juga membedakan antara teks subjektif dan objektif. Teks subjektif mengandung opini, evaluasi, atau ekspresi perasaan. Teks objektif berisi fakta atau informasi deskriptif tanpa sikap pribadi yang jelas. Pemisahan ini penting karena sistem analisis sentimen biasanya bekerja

lebih efektif pada teks yang bersifat subjektif. Identifikasi subjektivitas sering menjadi tahap awal sebelum proses klasifikasi sentimen dilakukan.

Konsep polarity merujuk pada arah sentimen, apakah positif atau negatif. Konsep intensity merujuk pada tingkat kekuatan sentimen tersebut. Dua teks dapat sama-sama bernilai positif, tetapi memiliki intensitas yang berbeda. Kalimat "produk ini baik" memiliki intensitas lebih rendah dibandingkan "produk ini sangat luar biasa." Perbedaan intensitas ini memengaruhi interpretasi hasil analisis, terutama dalam kajian pemasaran atau kebijakan publik.

Ruang lingkup analisis sentimen juga dapat dibedakan berdasarkan level analisis. Setiap level memiliki fokus dan kompleksitas yang berbeda. Perbandingan ketiga level tersebut dapat dilihat pada tabel berikut

Level Analisis	Unit yang Dianalisis	Contoh Kasus	Kompleksitas
Document-level	Seluruh dokumen	Ulasan produk secara keseluruhan	Rendah hingga sedang
Sentence-level	Kalimat tunggal	Komentar pendek di media sosial	Sedang
Aspect-level	Aspek atau fitur tertentu	Penilaian terhadap harga, kualitas, atau layanan	Tinggi

Pada level dokumen, sistem menentukan sentimen keseluruhan dari satu teks utuh. Pendekatan ini cocok untuk ulasan panjang seperti review film atau produk. Pada level kalimat, sistem menilai setiap kalimat secara terpisah. Pendekatan ini lebih sensitif terhadap variasi opini dalam satu dokumen. Pada level aspek, sistem mengidentifikasi bagian spesifik dari objek yang dievaluasi. Pendekatan ini lebih rinci karena satu teks dapat mengandung sentimen berbeda terhadap aspek yang berbeda.

.

C. SEJARAH PERKEMBANGAN ANALISIS SENTIMEN

Perkembangan analisis sentimen tidak terjadi secara tiba-tiba. Bidang ini tumbuh seiring kemajuan pemrosesan bahasa alami dan pembelajaran mesin. Setiap era membawa pendekatan baru dan mengubah cara peneliti memahami opini dalam teks.

Pada tahap awal, peneliti mengandalkan rule-based system. Pendekatan ini menggunakan aturan linguistik yang dirancang secara manual. Sistem mengidentifikasi kata positif dan negatif berdasarkan daftar tertentu. Peneliti juga menyusun pola tata bahasa untuk menangkap struktur kalimat sederhana. Pendekatan ini cukup efektif untuk domain terbatas, tetapi sulit beradaptasi

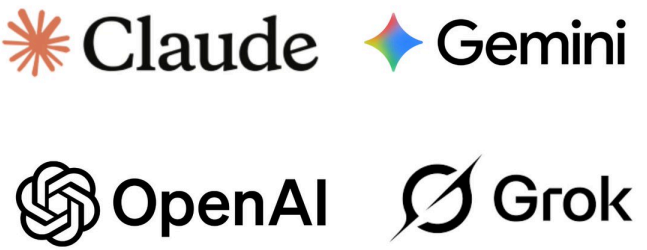
dengan variasi bahasa. Setiap perubahan konteks menuntut aturan baru. Proses ini memakan waktu dan sulit diskalakan.

Perkembangan berikutnya melahirkan lexicon-based approach. Pendekatan ini menggunakan kamus sentimen yang berisi daftar kata dan skor polaritas. Kamus seperti SentiWordNet banyak digunakan dalam penelitian awal tahun 2000-an. Sistem menghitung jumlah kata positif dan negatif dalam teks untuk menentukan kecenderungan sentimen. Pendekatan ini lebih fleksibel dibandingkan rule-based system, tetapi tetap bergantung pada kualitas dan cakupan kamus. Bahasa informal dan istilah baru sering tidak tercakup dalam leksikon yang tersedia.

Memasuki akhir 2000-an, machine learning mulai mendominasi penelitian analisis sentimen. Peneliti melatih model seperti Naive Bayes dan Support Vector Machine menggunakan data berlabel. Model tidak lagi bergantung pada aturan tetap, tetapi belajar dari pola dalam data. Pendekatan ini meningkatkan akurasi dan kemampuan generalisasi. Namun performa model sangat dipengaruhi oleh kualitas fitur dan jumlah data pelatihan. Representasi teks seperti Bag of Words dan TF-IDF menjadi komponen penting pada era ini.

Sekitar tahun 2014, deep learning membawa perubahan besar dalam analisis sentimen. Model seperti Convolutional Neural Network dan Long Short-Term Memory mampu menangkap hubungan kata secara lebih kompleks. Model ini tidak hanya melihat frekuensi kata, tetapi juga urutan dan konteksnya. Deep learning mengurangi ketergantungan pada rekayasa fitur manual. Namun model ini membutuhkan data besar dan sumber daya komputasi yang tinggi. Tidak semua institusi memiliki akses pada infrastruktur tersebut.

Perubahan berikutnya terjadi ketika transformer diperkenalkan pada tahun 2017 melalui arsitektur yang dikenal sebagai Transformer. Model ini mengandalkan mekanisme perhatian untuk memahami konteks dalam kalimat. Kemudian muncul model seperti BERT pada tahun 2018 yang menggunakan pretraining skala besar. Large language model memperluas kemampuan analisis sentimen dengan pemahaman konteks yang lebih dalam.



Logo Aplikasi Generative AI Berbasis Transformer

Model ini mampu menangani kalimat kompleks, ironi, dan variasi bahasa lintas domain. Meski demikian, isu bias data dan transparansi model masih menjadi perhatian dalam penelitian terkini. Perkembangan tersebut dapat diringkas dalam tabel berikut.

Periode	Era	Karakteristik Utama	Keterbatasan
1990–2004	Rule-based system	Aturan linguistik manual	Sulit diskalakan dan adaptif
2004–2010	Lexicon-based approach	Kamus sentimen dan skor polaritas	Bergantung pada kualitas leksikon
2008–2014	Machine learning	Model klasifikasi berbasis data	Butuh rekayasa fitur manual
2014–2018	Deep learning	CNN, RNN, LSTM untuk teks	Butuh data besar dan komputasi tinggi
2018–Saat ini	Transformer dan LLM	BERT, GPT, pretraining skala besar	Isu bias dan interpretabilitas

Sejarah ini menunjukkan bahwa analisis sentimen berkembang mengikuti kemajuan teknologi NLP. Setiap era membawa solusi baru, tetapi juga menghadirkan tantangan baru. Perubahan tersebut terus berlangsung seiring peningkatan kapasitas komputasi dan ketersediaan data global.

D. KONSEP DASAR DALAM ANALISIS SENTIMEN

Analisis sentimen tidak hanya berkaitan dengan klasifikasi positif atau negatif. Bidang ini berdiri di atas sejumlah konsep dasar yang memengaruhi hasil interpretasi teks. Pemahaman terhadap konsep tersebut membantu peneliti menghindari kesalahan dalam pemodelan dan evaluasi.

Sentimen berbeda dari emosi, meskipun keduanya sering tumpang tindih. Sentimen merujuk pada sikap umum terhadap suatu objek. Emosi merujuk pada kondisi afektif yang lebih spesifik seperti marah, senang, takut, atau sedih. Sebuah ulasan dapat bernilai positif tanpa menunjukkan emosi yang kuat. Sebaliknya, sebuah komentar dapat mengandung emosi intens tetapi tidak jelas arahnya terhadap objek tertentu. Perbedaan ini memengaruhi pendekatan klasifikasi yang digunakan dalam penelitian.

Dalam setiap opini terdapat tiga elemen utama, yaitu opinion holder, opinion target, dan orientation. Opinion holder adalah pihak yang

menyampaikan opini. Opinion target adalah objek atau aspek yang dinilai. Orientation merujuk pada arah sentimen, apakah positif atau negatif. Ketiga elemen ini membentuk struktur dasar analisis opini. Tanpa identifikasi target yang jelas, sistem dapat salah menafsirkan arah sentimen.

Konteks memegang peran penting dalam interpretasi teks. Satu kata dapat memiliki makna berbeda tergantung pada kalimat dan situasi. Kata “panas” dapat menggambarkan cuaca atau menggambarkan situasi politik yang tegang. Model yang tidak mempertimbangkan konteks sering menghasilkan klasifikasi yang keliru. Ambiguitas bahasa menjadi tantangan serius dalam data media sosial yang pendek dan padat makna.

Sarkasme dan ironi menambah kompleksitas analisis. Sebuah kalimat seperti “pelayanannya cepat sekali, sampai dua jam menunggu” secara literal tampak positif, tetapi sebenarnya mengandung kritik. Sistem berbasis kata kunci sering gagal mengenali pola semacam ini. Model yang lebih canggih mencoba memanfaatkan konteks yang lebih luas atau informasi tambahan seperti tanda baca dan pola historis pengguna. Namun pengenalan sarkasme masih menjadi masalah terbuka dalam banyak penelitian.

Bahasa informal dan slang juga memengaruhi akurasi sistem. Pengguna media sosial sering menyingkat kata atau mencampur bahasa. Kata seperti “mantul” atau “gaje” tidak selalu tercantum dalam kamus standar. Variasi ejaan dan penggunaan emoji memperumit proses tokenisasi dan normalisasi. Peneliti perlu melakukan pra-pemrosesan yang cermat untuk mengurangi gangguan ini.

Konsep-konsep dasar tersebut menunjukkan bahwa analisis sentimen bukan sekadar tugas klasifikasi sederhana. Bahasa manusia bersifat dinamis dan kontekstual. Setiap model harus mempertimbangkan struktur opini, konteks penggunaan, serta karakteristik bahasa yang digunakan. Pemahaman konseptual ini menjadi landasan sebelum membahas pendekatan metodologis pada bagian berikutnya.

E. KOMPONEN UTAMA SISTEM ANALISIS SENTIMEN

Sistem analisis sentimen bekerja melalui serangkaian tahapan yang saling terhubung. Setiap tahap memiliki fungsi yang jelas dan menentukan kualitas hasil akhir. Kegagalan pada satu tahap akan memengaruhi keseluruhan sistem. Oleh karena itu, perancangan arsitektur perlu dilakukan secara sistematis.

Tahap pertama adalah akuisisi data. Peneliti mengumpulkan teks dari media sosial, ulasan produk, forum diskusi, atau survei daring. Proses ini dapat menggunakan API resmi atau teknik web scraping sesuai aturan etika. Data yang terkumpul biasanya masih mentah dan belum siap dianalisis. Jumlah dan kualitas data sangat menentukan performa model pada tahap berikutnya.

Tahap kedua adalah pra-pemrosesan. Sistem membersihkan teks dari karakter yang tidak relevan. Proses ini mencakup case folding, tokenisasi,

penghapusan stopword, dan normalisasi kata tidak baku. Pada data bahasa Indonesia, proses stemming sering menggunakan algoritma khusus seperti Nazief dan Adriani. Tahap ini bertujuan mengurangi noise dan menyederhanakan struktur teks tanpa menghilangkan makna utama.

Tahap ketiga adalah ekstraksi fitur. Sistem mengubah teks menjadi representasi numerik agar dapat diproses oleh algoritma pembelajaran mesin. Representasi dapat berupa Bag of Words, TF-IDF, atau word embedding. Pilihan metode bergantung pada jenis model yang digunakan. Representasi yang tepat membantu model menangkap pola sentimen secara lebih akurat.

Tahap keempat adalah pemodelan. Pada tahap ini, algoritma dilatih menggunakan data berlabel. Model seperti Logistic Regression, Support Vector Machine, atau LSTM mempelajari pola hubungan antara fitur dan label sentimen. Proses pelatihan melibatkan pembagian data menjadi data latih dan data uji. Penyesuaian parameter dilakukan untuk memperoleh kinerja optimal.

Tahap kelima adalah evaluasi. Sistem mengukur kinerja model menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Evaluasi membantu peneliti memahami kekuatan dan kelemahan model. Analisis kesalahan sering dilakukan untuk melihat jenis kesalahan yang dominan. Hasil evaluasi menjadi dasar untuk perbaikan sistem.

Tahap terakhir adalah deployment. Model yang telah diuji kemudian diintegrasikan ke dalam sistem aplikasi. Implementasi dapat berbentuk dashboard analitik, API layanan, atau aplikasi web interaktif. Pada tahap ini, sistem harus mampu menangani data baru secara konsisten. Pemantauan berkala diperlukan untuk menjaga performa model dalam jangka panjang. Alur umum sistem analisis sentimen dapat digambarkan sebagai berikut.

Diagram Alur Analisis Sentimen

Diagram tersebut menunjukkan aliran proses yang bersifat linier, tetapi dalam praktiknya sering terjadi iterasi. Jika hasil evaluasi belum memadai, peneliti dapat kembali ke tahap pemodelan atau bahkan ke tahap pra-pemrosesan. Proses ini berlangsung hingga sistem mencapai performa yang dapat diterima dalam konteks penelitian atau kebutuhan industri.

F. JENIS PENDEKATAN DALAM ANALISIS SENTIMEN

Analisis sentimen berkembang melalui beberapa pendekatan utama yang memiliki karakteristik berbeda. Setiap pendekatan menawarkan cara tersendiri dalam memahami opini dalam teks. Bagian ini memberikan gambaran umum sebelum pembahasan teknis pada bab berikutnya.

Pendekatan rule-based mengandalkan aturan linguistik yang dirancang secara manual. Peneliti menyusun pola kalimat dan daftar kata yang menunjukkan sentimen tertentu. Sistem kemudian mencocokkan teks dengan aturan tersebut untuk menentukan klasifikasi. Pendekatan ini mudah dipahami dan transparan. Namun sistem sulit beradaptasi ketika bahasa berubah atau konteks berbeda.

Pendekatan lexicon-based menggunakan kamus sentimen yang berisi kata dan skor polaritas. Sistem menghitung jumlah kata positif dan negatif dalam teks. Hasil perhitungan tersebut menentukan kecenderungan sentimen dokumen. Pendekatan ini lebih fleksibel dibandingkan rule-based karena tidak memerlukan banyak aturan tata bahasa. Meski demikian, kualitas hasil sangat bergantung pada kelengkapan dan akurasi kamus yang digunakan.

Pendekatan machine learning-based memanfaatkan data berlabel untuk melatih model klasifikasi. Algoritma seperti Naive Bayes, Logistic Regression, dan Support Vector Machine mempelajari pola dari fitur numerik. Model tidak lagi bergantung pada aturan tetap, tetapi belajar dari contoh nyata. Pendekatan ini biasanya menghasilkan akurasi yang lebih tinggi dibandingkan metode berbasis kamus. Namun model memerlukan data pelatihan yang cukup dan proses validasi yang ketat.

Pendekatan deep learning-based menggunakan jaringan saraf dengan banyak lapisan. Model seperti LSTM dan CNN mampu menangkap hubungan kata secara lebih kompleks. Pendekatan ini sering memanfaatkan word embedding untuk memahami konteks. Dalam beberapa kasus, model transformer seperti BERT digunakan untuk meningkatkan performa. Deep learning mengurangi kebutuhan rekayasa fitur manual, tetapi memerlukan sumber daya komputasi yang besar.

Pendekatan hybrid menggabungkan dua atau lebih metode dalam satu sistem. Peneliti dapat memadukan kamus sentimen dengan model pembelajaran mesin. Sistem dapat menggunakan aturan linguistik untuk tahap awal, lalu memanfaatkan model statistik untuk klasifikasi akhir. Pendekatan ini mencoba memanfaatkan keunggulan masing-masing metode. Namun perancangannya lebih kompleks dan memerlukan pengujian yang cermat.

Gambaran umum perbandingan pendekatan tersebut dapat dilihat pada tabel berikut.

Pendekatan	Sumber Pengetahuan	Kelebihan	Keterbatasan
Rule-based	Aturan linguistik manual	Transparan dan mudah dipahami	Sulit adaptif terhadap domain baru
Lexicon-based	Kamus sentimen	Implementasi relatif sederhana	Bergantung pada kualitas leksikon
Machine learning	Data berlabel	Akurasi lebih tinggi dan adaptif	Membutuhkan data pelatihan
Deep learning	Jaringan saraf dan embedding	Menangkap konteks kompleks	Butuh komputasi dan data besar
Hybrid	Kombinasi beberapa metode	Lebih fleksibel dan seimbang	Desain sistem lebih rumit

Kelima pendekatan tersebut menunjukkan bahwa tidak ada metode yang sepenuhnya unggul dalam semua situasi. Pemilihan pendekatan bergantung pada tujuan penelitian, ketersediaan data, serta sumber daya yang dimiliki. Bab-bab berikutnya akan membahas masing-masing pendekatan secara lebih rinci dan aplikatif.

BAB 2. KONSEP DASAR DAN LANDASAN TEORETIS ANALISIS SENTIMEN

Yogiek Indra Kurniawan, S.T., M.T.

Analisis sentimen tidak berdiri sendiri sebagai teknik klasifikasi sederhana. Kajian ini berakar pada pemrosesan bahasa alami, linguistik, dan teori psikologi tentang sikap dan emosi. Pemahaman atas dasar teoretis membantu peneliti merancang model yang lebih tepat dan terukur.

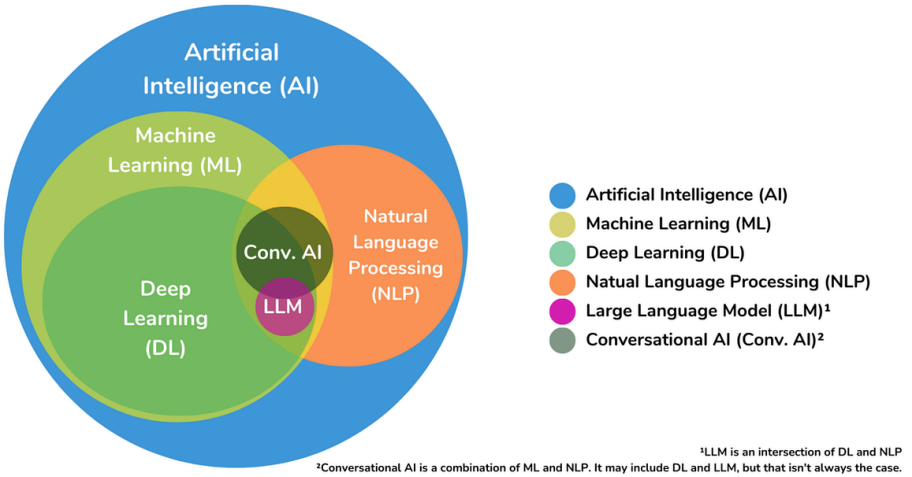
A. NATURAL LANGUAGE PROCESSING DALAM KERANGKA KECERDASAN BUATAN

Natural Language Processing merupakan cabang kecerdasan buatan yang mempelajari interaksi antara komputer dan bahasa manusia. Bidang ini bertujuan membuat mesin mampu memahami, memproses, dan menghasilkan teks secara otomatis. NLP bekerja pada data tidak terstruktur yang berasal dari percakapan, dokumen, atau media sosial. Tantangan utamanya terletak pada kompleksitas dan variasi bahasa alami.

Ruang lingkup NLP mencakup berbagai tugas pemrosesan teks. Tugas tersebut meliputi tokenisasi, pelabelan kelas kata, analisis struktur kalimat, dan klasifikasi dokumen. Tokenisasi memecah teks menjadi unit kata atau frasa. Tagging memberikan label gramatikal pada setiap kata. Parsing menganalisis struktur sintaksis dalam kalimat. Klasifikasi menentukan kategori tertentu berdasarkan isi teks.

Perkembangan NLP menunjukkan perubahan pendekatan yang cukup signifikan. Pada tahap awal, sistem berbasis aturan mendominasi penelitian. Peneliti merancang aturan linguistik secara manual untuk menangani pola kalimat tertentu. Pendekatan ini efektif untuk domain terbatas, tetapi sulit menangani variasi bahasa yang luas. Seiring meningkatnya ketersediaan data

digital pada awal 2000-an, pendekatan berbasis pembelajaran mesin mulai berkembang.



Ilustrasi Hubungan NLP dengan AI dan LLM
(https://medium.com/@sudhanshu.bhargav/genai-101-ai-ml-llms-explained-2e324aa2a0b0)

Model statistik seperti Hidden Markov Model dan Support Vector Machine mulai digunakan dalam berbagai tugas NLP. Model ini belajar dari data berlabel dan tidak lagi bergantung sepenuhnya pada aturan manual. Perkembangan berikutnya ditandai dengan munculnya model neural yang memanfaatkan jaringan saraf tiruan. Model neural mampu menangkap hubungan nonlinier yang lebih kompleks dalam teks. Transformasi ini mengubah cara mesin memahami konteks dan makna.

Analisis sentimen menempati posisi khusus dalam kerangka NLP. Tugas ini termasuk dalam kategori klasifikasi teks. Sistem tidak hanya mengenali struktur bahasa, tetapi juga mengidentifikasi sikap yang terkandung dalam teks. Analisis sentimen memanfaatkan hasil tokenisasi, tagging, dan representasi teks untuk menentukan orientasi opini. Dengan demikian, keberhasilan analisis sentimen sangat bergantung pada kualitas proses NLP sebelumnya.

Hubungan NLP dengan machine learning dan deep learning bersifat erat dan saling mendukung. Machine learning menyediakan metode untuk mempelajari pola dari data teks. Deep learning memperluas kemampuan tersebut dengan model berbasis jaringan saraf dalam. NLP menyediakan struktur dan representasi bahasa, sementara machine learning menyediakan mekanisme pembelajaran. Integrasi ketiganya membentuk fondasi metodologis dalam penelitian analisis sentimen modern.

B. LINGUISTIK KOMPUTASIONAL

Linguistik komputasional merupakan bidang yang menggabungkan ilmu bahasa dan komputasi. Bidang ini mengkaji bagaimana struktur bahasa dapat dimodelkan secara formal dan diproses oleh mesin. Fokusnya tidak hanya pada pemrograman, tetapi juga pada teori bahasa yang mendasari pemrosesan tersebut. Dalam analisis sentimen, linguistik komputasional menyediakan kerangka untuk memahami bagaimana makna dibentuk dalam teks.

Bahasa memiliki beberapa lapisan struktur yang saling berkaitan. Setiap lapisan berkontribusi pada pembentukan makna. Lapisan tersebut dapat diringkas dalam tabel berikut.

Lapisan	Fokus Kajian	Relevansi dalam Analisis Sentimen
Fonologi	Sistem bunyi	Terbatas pada teks lisan atau transkripsi
Morfologi	Struktur kata	Penting untuk stemming dan lemmatization
Sintaksis	Struktur kalimat	Menentukan hubungan antar kata
Semantik	Makna kata dan kalimat	Menentukan orientasi opini
Pragmatik	Makna dalam konteks	Menginterpretasi maksud implisit

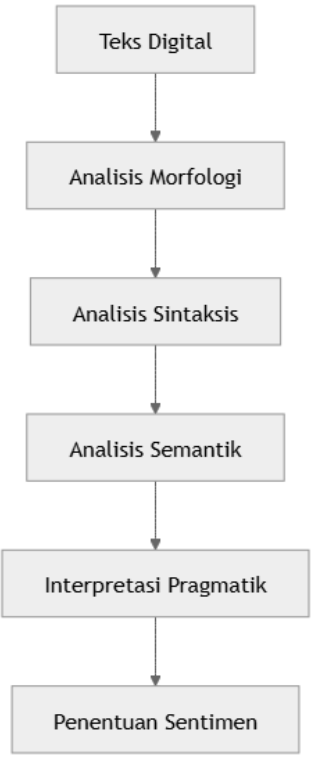
Fonologi jarang menjadi fokus utama dalam analisis sentimen berbasis teks. Namun morfologi berperan penting dalam normalisasi kata. Sintaksis membantu sistem memahami hubungan subjek dan objek dalam kalimat. Semantik menentukan makna yang lebih dalam dari kata atau frasa. Pragmatik membantu menjelaskan maksud yang tidak tertulis secara eksplisit.

Sintaksis dan semantik memiliki peran sentral dalam interpretasi opini. Struktur sintaksis menentukan siapa yang melakukan tindakan dan terhadap apa tindakan itu diarahkan. Kesalahan dalam memahami struktur dapat menghasilkan klasifikasi yang keliru. Sebagai contoh, kalimat dengan negasi dapat mengubah makna secara drastis. Semantik membantu sistem memahami bahwa kata tertentu memiliki muatan evaluatif tertentu dalam konteks tertentu.

Ambiguitas menjadi tantangan utama dalam linguistik komputasional. Ambiguitas leksikal terjadi ketika satu kata memiliki lebih dari satu makna. Kata “ringan” dapat merujuk pada berat fisik atau tingkat kesulitan. Ambiguitas struktural terjadi ketika susunan kalimat memungkinkan lebih dari satu

interpretasi. Model yang tidak mempertimbangkan konteks sering memilih makna yang tidak tepat.

Hubungan antara struktur bahasa dan interpretasi opini dapat digambarkan sebagai berikut.



Ilustrasi Hubungan Struktur Bahasa dan Interpretasi Opini

Diagram tersebut menunjukkan bahwa pemahaman sentimen tidak muncul secara langsung dari kata tunggal. Sistem perlu melalui beberapa lapisan analisis sebelum menentukan orientasi opini.

Pragmatik menjadi semakin penting dalam teks digital. Pengguna media sosial sering menggunakan ironi, singkatan, atau referensi budaya tertentu. Makna tidak selalu tersurat dalam kata yang digunakan. Konteks percakapan, situasi sosial, dan bahkan tren internet memengaruhi interpretasi. Tanpa pemahaman pragmatik, sistem dapat gagal mengenali maksud sebenarnya dari suatu pernyataan.

Linguistik komputasional membantu menjembatani teori bahasa dan implementasi algoritmik. Dengan memahami struktur dan konteks bahasa, peneliti dapat merancang sistem analisis sentimen yang lebih akurat dan adaptif terhadap dinamika komunikasi digital.

C. TEORI SENTIMEN DAN EMOSI

Kajian analisis sentimen tidak dapat dilepaskan dari teori tentang sikap dan emosi dalam psikologi. Sentimen dan emosi sering digunakan secara bergantian dalam percakapan sehari-hari. Namun dalam kajian ilmiah, keduanya memiliki perbedaan yang cukup jelas. Pemahaman perbedaan ini membantu peneliti menentukan batas analisis yang tepat.

Sentimen merujuk pada sikap evaluatif terhadap suatu objek atau isu. Sikap ini cenderung stabil dalam konteks tertentu. Emosi merujuk pada keadaan afektif yang lebih intens dan sering bersifat sementara. Seseorang dapat memiliki sentimen positif terhadap sebuah merek, tetapi tetap merasa marah karena layanan yang lambat pada hari tertentu. Dengan demikian, sentimen berhubungan dengan penilaian, sedangkan emosi berhubungan dengan pengalaman perasaan.

Positif, Negatif dan Netral Merupakan Sentimen Paling Umum

Dalam praktik analisis sentimen, klasifikasi paling umum mencakup tiga kategori dasar, yaitu positif, negatif, dan netral. Kategori positif menunjukkan dukungan atau kepuasan. Kategori negatif menunjukkan penolakan atau ketidakpuasan. Kategori netral mencerminkan pernyataan informatif tanpa muatan evaluatif yang jelas. Meskipun pembagian ini sederhana, penerapannya sering menghadapi tantangan karena batas antar kategori tidak selalu tegas.

Psikologi emosi menawarkan beberapa model dasar yang membantu memahami variasi ekspresi afektif. Salah satu pendekatan awal mengelompokkan emosi menjadi beberapa jenis utama seperti marah, takut, senang, dan sedih. Pendekatan lain memandang emosi dalam dimensi valensi dan arousal. Valensi merujuk pada arah positif atau negatif. Arousal merujuk pada tingkat intensitas atau aktivasi emosional. Model dimensi ini sering digunakan dalam analisis sentimen berbasis skala kontinu.

Hubungan antara evaluasi kognitif dan ekspresi bahasa menjadi aspek penting dalam kajian ini. Teori appraisal menjelaskan bahwa emosi muncul dari penilaian individu terhadap suatu peristiwa. Penilaian tersebut kemudian diwujudkan dalam pilihan kata dan struktur kalimat. Kalimat yang mengandung kata evaluatif biasanya mencerminkan proses kognitif tertentu. Oleh karena itu, analisis sentimen tidak hanya membaca kata, tetapi juga menafsirkan hasil evaluasi mental yang tercermin dalam bahasa.

Representasi emosi dalam teks memiliki karakteristik tersendiri. Penulis dapat menyampaikan emosi secara eksplisit melalui kata sifat atau kata kerja tertentu. Penulis juga dapat menyampaikannya secara implisit melalui metafora atau ironi. Dalam teks digital, emoji dan tanda baca sering memperkuat ekspresi emosi. Tantangan muncul ketika emosi disampaikan secara tidak langsung atau bertentangan dengan makna literal kalimat.

Pemahaman teori sentimen dan emosi membantu memperluas cakupan analisis. Sistem tidak hanya menentukan arah sentimen, tetapi juga mempertimbangkan intensitas dan variasi ekspresi. Landasan teoretis ini penting agar analisis sentimen tidak terjebak pada klasifikasi dangkal tanpa memahami dinamika psikologis yang melatarbelakanginya.

D. SUBJECTIVITY DAN OBJECTIVITY

Pembahasan subjektivitas dan objektivitas menjadi tahap penting sebelum klasifikasi sentimen dilakukan. Tidak semua teks mengandung opini. Sebagian teks hanya menyampaikan informasi faktual tanpa evaluasi pribadi. Sistem analisis sentimen perlu membedakan keduanya agar tidak menghasilkan interpretasi yang keliru.

Teks subjektif memuat sikap, penilaian, atau ekspresi perasaan penulis. Teks objektif menyampaikan fakta yang dapat diverifikasi tanpa muatan evaluatif. Kalimat seperti “produk ini sangat nyaman digunakan” bersifat subjektif karena mengandung penilaian pribadi. Kalimat “produk ini dirilis pada tahun 2023” bersifat objektif karena menyampaikan fakta. Perbedaan ini dapat dirangkum dalam tabel berikut.

Aspek	Subjektif	Objektif
Isi utama	Opini atau evaluasi	Fakta atau informasi
Bahasa	Mengandung kata evaluatif	Bahasa deskriptif netral

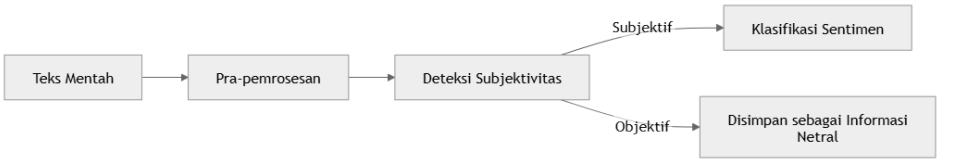
Verifikasi	Bergantung pada persepsi	Dapat diuji kebenarannya
Relevansi untuk analisis sentimen	Sangat tinggi	Terbatas

Identifikasi kalimat opini dan fakta sering dilakukan sebelum tahap klasifikasi sentimen. Sistem dapat menggunakan daftar kata evaluatif atau model pembelajaran mesin untuk memisahkan keduanya. Proses ini membantu mengurangi noise pada data. Jika teks objektif langsung diklasifikasikan sebagai positif atau negatif, hasilnya dapat menyesatkan. Oleh karena itu, klasifikasi subjektivitas sering menjadi tahap awal dalam pipeline analisis.

Peran subjektivitas dalam analisis sentimen bersifat mendasar. Sentimen pada dasarnya merupakan bentuk evaluasi subjektif terhadap suatu objek. Tanpa elemen subjektif, tidak ada orientasi sentimen yang dapat diidentifikasi. Pemisahan antara fakta dan opini membantu sistem fokus pada bagian teks yang relevan. Pendekatan ini juga meningkatkan akurasi model dalam banyak penelitian empiris.

Namun klasifikasi subjektivitas tidak selalu mudah. Beberapa kalimat mengandung campuran fakta dan opini. Kalimat seperti "produk ini memiliki baterai 5000 mAh yang sangat awet" memuat fakta teknis dan penilaian sekaligus. Sistem perlu mengenali bagian mana yang bersifat evaluatif. Tantangan juga muncul ketika opini disampaikan secara implisit tanpa kata sifat yang jelas. Model harus mampu menangkap pola yang lebih halus dalam struktur kalimat.

Alur identifikasi subjektivitas dalam sistem dapat digambarkan sebagai berikut.



Contoh kasus dalam ulasan produk menunjukkan pentingnya tahap ini. Ulasan sering menggabungkan spesifikasi teknis dan pengalaman pengguna. Bagian spesifikasi cenderung objektif, sedangkan pengalaman bersifat subjektif. Dalam kebijakan publik, laporan resmi biasanya objektif, tetapi komentar masyarakat bersifat subjektif. Sistem yang tidak membedakan keduanya dapat mencampur fakta dengan opini.

Pemahaman terhadap subjektivitas membantu memperjelas batas analisis sentimen. Pendekatan ini membuat sistem lebih selektif dan kontekstual.

Dengan dasar ini, model dapat menghasilkan interpretasi yang lebih akurat dan relevan terhadap tujuan penelitian.

E. POLARITY DAN INTENSITY

Polarity dan intensity merupakan dua konsep utama dalam pengukuran sentimen. Keduanya menentukan bagaimana sistem menilai kecenderungan opini dalam teks. Tanpa pemahaman yang jelas, hasil klasifikasi dapat kehilangan makna analitis.

Polarity merujuk pada arah sentimen terhadap suatu objek. Arah ini biasanya dibagi menjadi positif dan negatif. Beberapa penelitian menambahkan kategori netral untuk teks tanpa evaluasi yang jelas. Polarity membantu menjawab pertanyaan dasar apakah penulis mendukung atau menolak suatu objek. Konsep ini menjadi fondasi dalam banyak model klasifikasi biner dan multikelas.

Intensity merujuk pada tingkat kekuatan sentimen tersebut. Dua kalimat dapat memiliki polarity yang sama tetapi berbeda dalam intensitas. Kalimat "layanan ini baik" menunjukkan sentimen positif dengan intensitas rendah. Kalimat "layanan ini luar biasa dan sangat memuaskan" menunjukkan intensitas yang lebih tinggi. Perbedaan ini penting dalam analisis pemasaran dan pemantauan opini publik.

Perbedaan antara polarity dan intensity dapat dirangkum dalam tabel berikut.

Aspek	Polarity	Intensity
Fokus	Arah sentimen	Kekuatan sentimen
Nilai umum	Positif, negatif, netral	Rendah, sedang, tinggi
Fungsi utama	Menentukan kategori	Mengukur derajat ekspresi
Pengaruh pada analisis	Menentukan kelas	Memperhalus interpretasi

Dalam praktiknya, skala sentimen dapat bersifat diskrit atau kontinu. Skala diskrit membagi sentimen ke dalam kategori tetap seperti tiga atau lima kelas. Skala kontinu menggunakan rentang nilai numerik, misalnya dari −1 hingga 1. Skala kontinu memungkinkan analisis yang lebih halus, tetapi

memerlukan metode anotasi yang konsisten. Pilihan skala bergantung pada tujuan penelitian dan ketersediaan data.

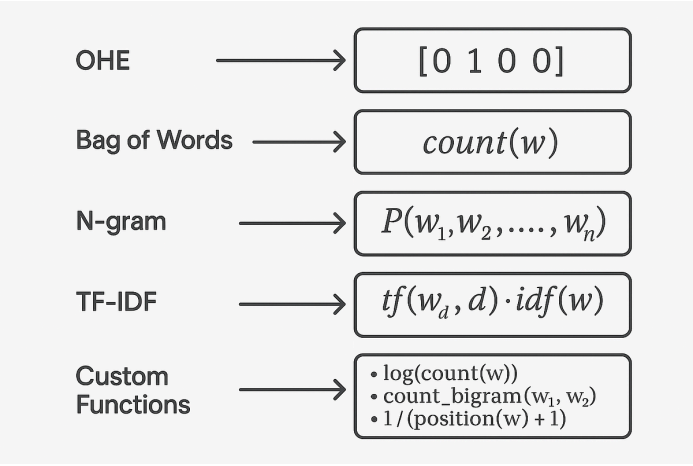
Sentiment scoring digunakan untuk memberikan nilai numerik pada teks. Nilai ini dapat dihitung berdasarkan kamus sentimen atau hasil prediksi model. Skema anotasi menjadi penting dalam proses pelabelan data latih. Anotator perlu memiliki pedoman yang jelas agar konsistensi terjaga. Perbedaan interpretasi antar anotator dapat menurunkan reliabilitas data.

Implikasi konsep ini terhadap model klasifikasi cukup signifikan. Model biner sederhana hanya memprediksi arah sentimen tanpa mempertimbangkan kekuatan ekspresi. Model regresi atau model multikelas dapat menangkap variasi intensitas secara lebih rinci. Peneliti perlu menyesuaikan desain model dengan kebutuhan analisis. Dalam studi opini publik, intensitas sering menjadi indikator perubahan sikap yang lebih sensitif dibandingkan kategori dasar.

Dengan memahami polarity dan intensity, analisis sentimen tidak berhenti pada label sederhana. Pendekatan ini membuka ruang untuk interpretasi yang lebih mendalam dan kontekstual.

F. TEORI REPRESENTASI TEKS

Representasi teks menjadi jembatan antara bahasa manusia dan model komputasional. Mesin tidak memahami kata seperti manusia memahami makna. Sistem membutuhkan bentuk numerik agar dapat memproses teks secara matematis. Oleh karena itu, teori representasi teks memiliki posisi sentral dalam analisis sentimen.



Ilustrasi Representasi Teks di NLP

(https://python.plainenglish.io/text-representation-in-nlp-0043941fd99d)

Secara umum, terdapat dua pendekatan utama dalam representasi teks, yaitu simbolik dan distribusional. Pendekatan simbolik memandang bahasa sebagai kumpulan simbol yang dapat dihitung dan diklasifikasikan. Setiap kata diperlakukan sebagai unit terpisah tanpa mempertimbangkan konteks luas. Pendekatan distribusional memandang makna kata berdasarkan konteks kemunculannya dalam korpus besar. Perbedaan keduanya dapat diringkas dalam tabel berikut.

Aspek	Simbolik	Distribusional
Dasar teori	Representasi eksplisit	Pola kemunculan dalam konteks
Hubungan antar kata	Terbatas	Dipelajari dari data
Sensitivitas konteks	Rendah	Lebih tinggi
Contoh metode	Bag of Words	Word embedding

Pendekatan distribusional berakar pada distribusional hypothesis dalam linguistik. Hipotesis ini menyatakan bahwa kata yang muncul dalam konteks yang mirip cenderung memiliki makna yang mirip. Prinsip ini menjadi dasar bagi model embedding modern. Dengan mempelajari pola kemunculan kata dalam jutaan kalimat, sistem dapat menangkap hubungan semantik secara implisit.

Bag of Words menjadi representasi awal yang banyak digunakan dalam penelitian awal analisis sentimen. Metode ini mengubah dokumen menjadi vektor berdasarkan frekuensi kata. Urutan kata diabaikan, sehingga struktur sintaksis tidak dipertimbangkan. Meskipun sederhana, metode ini cukup efektif untuk tugas klasifikasi dasar. Namun pendekatan ini menghasilkan vektor yang jarang dan berdimensi tinggi.

Perkembangan berikutnya menghadirkan representasi berbasis frekuensi dan bobot seperti Term Frequency dan TF-IDF. Pendekatan ini tidak hanya menghitung jumlah kemunculan kata, tetapi juga mempertimbangkan distribusi kata dalam seluruh korpus. Kata yang terlalu umum mendapat bobot lebih

rendah. Pendekatan ini membantu menonjolkan kata yang lebih informatif dalam dokumen.

Word embedding membawa perubahan signifikan dalam representasi teks. Metode seperti Word2Vec dan GloVe memetakan kata ke dalam ruang vektor berdimensi rendah. Vektor ini menangkap hubungan semantik antar kata. Kata dengan makna yang mirip cenderung berada pada jarak yang dekat dalam ruang vektor. Perkembangan terbaru menghadirkan embedding kontekstual yang menyesuaikan representasi berdasarkan kalimat. Pendekatan ini membantu sistem memahami perbedaan makna kata dalam konteks berbeda.

Hubungan antara teori representasi dan performa model sangat erat. Representasi sederhana sering memadai untuk dataset kecil dengan struktur yang relatif stabil. Representasi berbasis embedding biasanya meningkatkan akurasi pada data yang kompleks dan beragam. Namun peningkatan kompleksitas juga menuntut sumber daya komputasi yang lebih besar. Pemilihan representasi harus mempertimbangkan tujuan penelitian, ukuran data, dan jenis model yang digunakan.

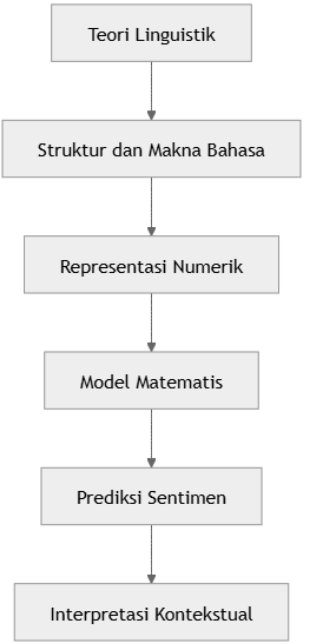
G. INTEGRASI KONSEP TEORETIS

Analisis sentimen tidak berkembang hanya dari inovasi algoritma. Bidang ini bertumpu pada integrasi antara teori linguistik dan komputasi. Linguistik menjelaskan bagaimana makna dibangun dalam bahasa. Komputasi menyediakan alat untuk memodelkan dan menghitung pola tersebut secara sistematis. Tanpa pemahaman bahasa, model hanya memproses simbol tanpa konteks.

Hubungan antara teori linguistik dan komputasi terlihat pada setiap tahap sistem. Analisis morfologi membantu proses normalisasi kata. Analisis sintaksis membantu memahami hubungan antar unsur kalimat. Analisis semantik membantu menentukan orientasi evaluatif. Komputasi kemudian mengubah hasil analisis tersebut menjadi representasi numerik yang dapat dipelajari oleh model. Integrasi ini membentuk fondasi yang saling melengkapi.

Subjektivitas dan konteks memperlihatkan interaksi yang lebih kompleks. Teks subjektif sering mengandung nuansa yang tidak eksplisit. Model matematis mencoba menangkap pola tersebut melalui fitur dan parameter. Namun konteks sosial dan budaya sering berada di luar jangkauan perhitungan numerik sederhana. Sistem perlu mempertimbangkan variasi domain dan latar penggunaan bahasa. Interaksi ini menunjukkan bahwa model tidak berdiri terpisah dari realitas sosial yang melingkupinya.

Hubungan antara unsur linguistik dan model matematis dapat digambarkan sebagai berikut.



Ilustrasi Teori Linguistik yang Digunakan Dalam NLP

Diagram tersebut menunjukkan bahwa prediksi sentimen merupakan hasil rangkaian proses konseptual dan komputasional. Interpretasi akhir tetap membutuhkan pemahaman konteks. Model dapat menghasilkan skor atau label, tetapi makna sosial dari skor tersebut perlu dianalisis secara kritis.

Pendekatan teknis yang berdiri tanpa dasar teoretis sering menghadapi keterbatasan. Model dapat mencapai akurasi tinggi pada dataset tertentu, tetapi gagal ketika diterapkan pada domain lain. Kegagalan ini sering terjadi karena model tidak memahami struktur dan dinamika bahasa secara mendalam. Tanpa landasan teori, pengembangan sistem cenderung bersifat coba dan salah. Pendekatan semacam ini sulit memberikan kontribusi ilmiah yang berkelanjutan.

Analisis sentimen pada akhirnya menempati posisi sebagai kajian interdisipliner. Bidang ini menggabungkan linguistik, psikologi, ilmu komputer, dan statistik. Peneliti perlu memahami aspek bahasa sekaligus aspek algoritmik. Pendekatan interdisipliner membantu menjembatani kesenjangan antara teori dan praktik. Dengan integrasi tersebut, analisis sentimen dapat berkembang sebagai bidang yang tidak hanya teknis, tetapi juga reflektif dan kontekstual.

BAB 3. PENDEKATAN DAN METODE ANALISIS SENTIMEN

Muhammad Adi Pratama, S.S., M.A.

A. KERANGKA UMUM PENDEKATAN ANALISIS SENTIMEN

Literatur analisis sentimen umumnya membagi pendekatan ke dalam dua kelompok besar, yaitu berbasis aturan dan berbasis data. Pendekatan berbasis aturan menekankan perancangan pola linguistik secara manual. Pendekatan berbasis data menekankan pembelajaran pola dari korpus berlabel. Perkembangan mutakhir kemudian melahirkan pendekatan hibrida dan berbasis jaringan saraf dalam.

Klasifikasi pendekatan dalam literatur dapat diringkas sebagai berikut.

Kategori Utama	Subkategori	Karakteristik Umum
Berbasis Aturan	Rule-based	Aturan linguistik eksplisit
Berbasis Leksikon	Lexicon-based	Kamus sentimen dan skor polaritas
Berbasis Data	Machine learning	Model statistik dengan data berlabel
Berbasis Neural	Deep learning	Representasi embedding dan jaringan saraf
Kombinasi	Hybrid	Integrasi aturan dan pembelajaran data

Pendekatan berbasis aturan bergantung pada daftar kata dan pola sintaksis yang dirancang secara eksplisit. Sistem menentukan sentimen berdasarkan kecocokan terhadap aturan tersebut. Pendekatan ini relatif mudah dipahami dan transparan. Namun sistem sulit beradaptasi ketika bahasa berubah atau domain berbeda.

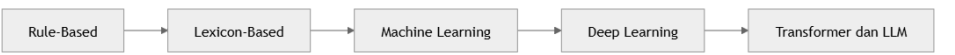
Pendekatan berbasis data menggunakan algoritma pembelajaran untuk menemukan pola secara otomatis. Model dilatih menggunakan data berlabel sehingga mampu melakukan generalisasi. Pendekatan ini lebih fleksibel dibandingkan sistem aturan. Namun performanya sangat bergantung pada kualitas dan jumlah data pelatihan.

Perbandingan keduanya dapat dilihat dalam tabel berikut.

Aspek	Berbasis Aturan	Berbasis Data
Sumber pengetahuan	Aturan manual	Data berlabel
Fleksibilitas	Rendah	Lebih tinggi
Kebutuhan data	Minimal	Tinggi
Transparansi	Tinggi	Tergantung model
Adaptasi lintas domain	Terbatas	Lebih adaptif

Evolusi metodologis dalam dua dekade terakhir menunjukkan pergeseran dari pendekatan manual menuju pendekatan berbasis pembelajaran mendalam. Pada awal 2000-an, rule-based dan lexicon-based mendominasi penelitian. Sekitar tahun 2010, machine learning mulai menggantikan pendekatan manual. Setelah 2014, deep learning memperkenalkan representasi kontekstual yang lebih kuat. Perkembangan transformer pada 2017 semakin memperluas kemampuan model dalam memahami konteks panjang.

Alur evolusi tersebut dapat digambarkan sebagai berikut.



Ilustrasi Evolusi Metode NLP

Pemilihan metode tidak dapat dilepaskan dari tujuan penelitian. Studi eksploratif dengan data terbatas dapat menggunakan pendekatan leksikon atau aturan. Penelitian dengan dataset besar dan kompleks cenderung menggunakan machine learning atau deep learning. Penelitian yang menuntut interpretabilitas tinggi sering memilih model yang lebih sederhana. Sebaliknya, penelitian yang menekankan akurasi maksimal mungkin memilih model neural yang lebih kompleks.

Kriteria pemilihan metode dapat diringkas dalam tabel berikut.

Tujuan Penelitian	Metode yang Disarankan
Data kecil dan domain spesifik	Rule-based atau lexicon-based
Dataset berlabel skala menengah	Machine learning klasik
Dataset besar dan kompleks	Deep learning
Kebutuhan interpretasi tinggi	Logistic Regression atau Decision Tree
Kebutuhan adaptasi lintas domain	Transformer atau model pralatih

Kerangka umum ini membantu peneliti memahami posisi setiap pendekatan. Tidak ada metode yang selalu unggul dalam semua situasi. Pemilihan metode harus mempertimbangkan konteks data, tujuan analisis, serta sumber daya yang tersedia.

B. RULE-BASED APPROACH

Pendekatan rule-based merupakan metode awal dalam analisis sentimen. Sistem bekerja berdasarkan seperangkat aturan linguistik yang dirancang secara manual. Aturan tersebut menentukan bagaimana kata atau pola kalimat dipetakan ke kategori sentimen tertentu. Pendekatan ini tidak memerlukan proses pelatihan berbasis data besar.

Konsep dasarnya cukup langsung. Sistem mencari kata kunci yang telah diberi label positif atau negatif. Jika jumlah kata positif lebih dominan, maka teks diklasifikasikan sebagai positif. Jika kata negatif lebih dominan, maka teks diklasifikasikan sebagai negatif. Pendekatan ini sering digunakan pada tahap awal penelitian atau pada domain yang sangat spesifik.

Penyusunan aturan dilakukan secara manual oleh peneliti atau ahli bahasa. Proses ini melibatkan identifikasi kata evaluatif, frasa tertentu, serta pola sintaksis yang relevan. Peneliti juga menentukan bobot atau prioritas aturan jika terjadi konflik. Penyusunan aturan membutuhkan pemahaman struktur bahasa dan konteks domain. Kesalahan dalam merancang aturan dapat menghasilkan bias sistematis.

Penggunaan pola sintaksis memperluas kemampuan sistem berbasis aturan. Sistem tidak hanya membaca kata tunggal, tetapi juga memperhatikan

hubungan antar kata. Sebagai contoh, pola adjektiva yang mengikuti kata benda sering menjadi indikator evaluasi. Pola seperti "sangat baik" atau "tidak memuaskan" dapat dirancang sebagai aturan khusus. Pendekatan ini membantu sistem mengenali ekspresi yang lebih kompleks dibandingkan pencocokan kata sederhana.

Penanganan negasi menjadi aspek penting dalam sistem berbasis aturan. Kata seperti "tidak" atau "bukan" dapat mengubah arah sentimen secara signifikan. Tanpa aturan khusus, sistem dapat salah mengklasifikasikan kalimat. Sebagai contoh, kalimat "produk ini tidak bagus" harus dikategorikan sebagai negatif meskipun mengandung kata "bagus." Sistem perlu mendeteksi keberadaan negasi dan membalik polaritas kata yang mengikutinya.

Alur kerja sederhana sistem rule-based dapat digambarkan sebagai berikut.

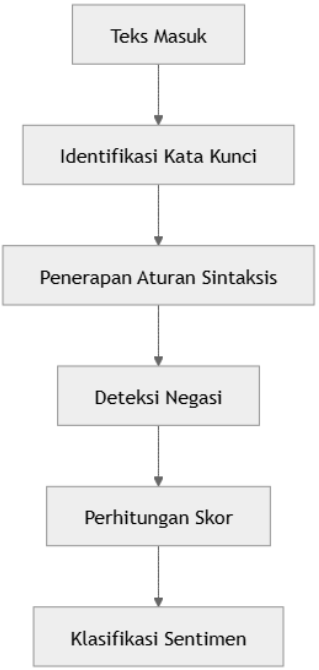


Diagram Alur Kerja Rule-Based Approach

Pendekatan rule-based memiliki beberapa kelebihan. Sistem relatif mudah dipahami dan transparan. Peneliti dapat menelusuri alasan di balik setiap keputusan klasifikasi. Metode ini juga tidak memerlukan dataset berlabel dalam jumlah besar. Pada domain sempit dengan kosakata terbatas, pendekatan ini dapat memberikan hasil yang cukup stabil.

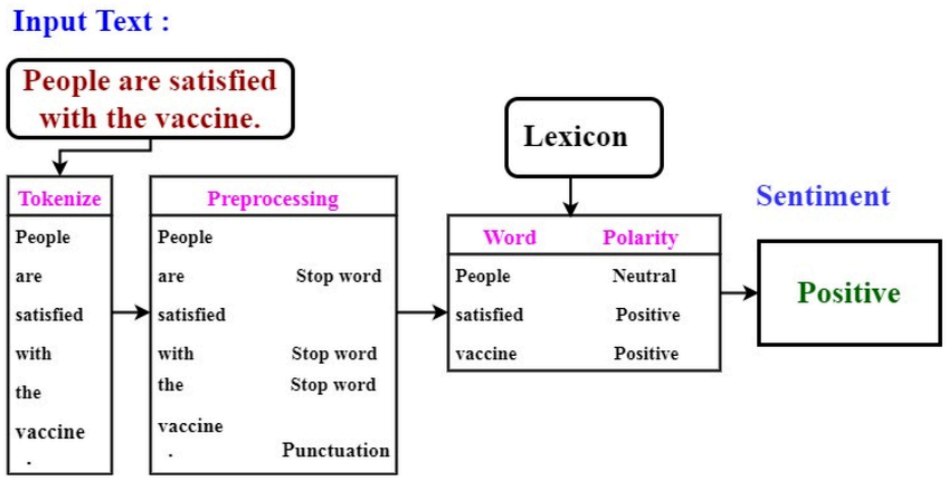
Namun pendekatan ini memiliki keterbatasan yang jelas. Sistem sulit beradaptasi terhadap variasi bahasa dan slang baru. Proses penyusunan aturan memakan waktu dan sulit diperluas ke domain lain. Sistem juga kesulitan

menangani sarkasme atau konteks yang kompleks. Ketika jumlah aturan bertambah banyak, konflik antar aturan dapat muncul dan mengurangi konsistensi hasil.

Contoh implementasi sederhana dapat dilakukan dengan langkah berikut. Peneliti menyusun daftar kata positif dan negatif. Sistem menghitung jumlah kemunculan masing-masing dalam satu dokumen. Jika skor positif lebih tinggi, dokumen diklasifikasikan sebagai positif. Jika skor negatif lebih tinggi, dokumen diklasifikasikan sebagai negatif. Meskipun sederhana, pendekatan ini menjadi fondasi bagi perkembangan metode yang lebih kompleks pada tahap berikutnya.

C. LEXICON-BASED APPROACH

Pendekatan lexicon-based berkembang dari sistem berbasis aturan yang lebih sederhana.



Proses NLP Berbasis Lexicon
(https://www.researchgate.net)

Metode ini menggunakan kamus sentimen yang berisi daftar kata beserta nilai polaritasnya. Setiap kata dalam kamus memiliki skor positif, negatif, atau netral. Sistem menghitung kecenderungan sentimen berdasarkan skor kata yang muncul dalam teks.

Prinsip dasarnya cukup jelas. Jika sebuah dokumen mengandung lebih banyak kata dengan skor positif, maka dokumen tersebut cenderung positif. Jika skor negatif lebih dominan, maka hasilnya negatif. Pendekatan ini tidak selalu memerlukan data latih berlabel. Sistem mengandalkan sumber leksikal yang telah disusun sebelumnya.

Penyusunan kamus sentimen menjadi tahap penting dalam metode ini. Peneliti dapat menggunakan kamus yang sudah tersedia atau menyusun kamus baru sesuai kebutuhan. Setiap entri kamus biasanya mencantumkan kata dan nilai polaritasnya. Nilai tersebut dapat berupa kategori diskrit atau skor numerik dalam rentang tertentu. Konsistensi penilaian menjadi faktor utama dalam kualitas kamus.

Terdapat dua jenis kamus yang umum digunakan, yaitu kamus umum dan kamus domain-spesifik. Kamus umum mencakup kosakata luas yang digunakan dalam berbagai konteks. Kamus domain-spesifik disusun untuk bidang tertentu seperti kesehatan, keuangan, atau pendidikan. Perbandingan keduanya dapat dilihat pada tabel berikut.

Aspek	Kamus Umum	Kamus Domain-Spesifik
Cakupan kosakata	Luas	Terbatas pada bidang tertentu
Adaptasi konteks	Lebih umum	Lebih tepat pada domain tertentu
Kemudahan penggunaan	Siap pakai	Perlu penyusunan tambahan
Akurasi lintas domain	Bervariasi	Tinggi pada domain target

Perhitungan skor sentimen dokumen biasanya dilakukan dengan menjumlahkan skor semua kata yang cocok dengan kamus. Beberapa metode menggunakan rata-rata skor untuk menyesuaikan panjang dokumen. Metode lain memberikan bobot tambahan pada kata yang memiliki intensifier seperti “sangat” atau “sekali.” Proses ini menghasilkan nilai akhir yang kemudian dipetakan ke kategori sentimen.

Meskipun sederhana, pendekatan ini menghadapi beberapa tantangan. Cakupan kosakata dalam kamus sering tidak lengkap. Bahasa informal, slang, dan istilah baru sering tidak tercantum dalam daftar. Perubahan makna kata dalam konteks tertentu juga dapat mengurangi akurasi. Kata yang bersifat positif dalam satu domain dapat bernilai netral atau negatif dalam domain lain.

Evaluasi performa metode lexicon-based biasanya menggunakan metrik klasifikasi standar seperti akurasi dan F1-score. Hasilnya sering memadai pada domain yang stabil dan formal. Namun pada data media sosial yang dinamis, performa dapat menurun. Pendekatan ini tetap relevan karena mudah diterapkan dan tidak membutuhkan proses pelatihan yang kompleks.

Dengan memahami prinsip dan keterbatasannya, pendekatan lexicon-based dapat digunakan secara tepat sesuai konteks penelitian. Metode

ini juga sering menjadi dasar dalam pengembangan pendekatan hibrida yang menggabungkan kamus dan pembelajaran mesin.

D. MACHINE LEARNING APPROACH

Pendekatan machine learning mengubah cara analisis sentimen dilakukan. Sistem tidak lagi bergantung pada aturan tetap atau kamus statis. Model belajar dari data berlabel untuk mengenali pola sentimen. Pendekatan ini termasuk dalam kerangka supervised learning karena proses pelatihan menggunakan pasangan teks dan label.

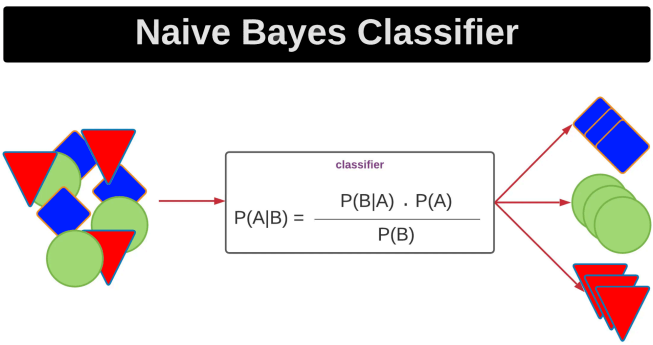
Dalam supervised learning, setiap dokumen memiliki label seperti positif atau negatif. Model mempelajari hubungan antara fitur teks dan label tersebut. Setelah pelatihan selesai, model digunakan untuk memprediksi sentimen pada data baru. Kualitas prediksi sangat dipengaruhi oleh jumlah dan keberagaman data latih.

Tahapan umum dalam pendekatan ini dapat diringkas sebagai berikut.

Alur NLP dengan Machine Learning

Tahap pertama adalah representasi fitur. Teks diubah menjadi vektor numerik menggunakan metode seperti Bag of Words atau TF-IDF. Tahap kedua adalah pelatihan model menggunakan data latih. Tahap ketiga adalah pengujian menggunakan data yang belum pernah dilihat model. Evaluasi dilakukan untuk mengukur akurasi dan metrik lain yang relevan.

Beberapa algoritma populer sering digunakan dalam klasifikasi sentimen. Naive Bayes merupakan model probabilistik yang sederhana dan efisien. Model ini bekerja dengan asumsi independensi antar fitur. Meskipun asumsi tersebut jarang terpenuhi secara penuh, model ini sering memberikan hasil yang kompetitif pada dataset teks.



Ilustrasi Naive Bayes

Logistic Regression menggunakan pendekatan probabilistik berbasis fungsi logistik. Model ini cocok untuk klasifikasi biner dan mudah diinterpretasikan. Support Vector Machine mencari hyperplane yang memisahkan kelas dengan margin maksimum. Algoritma ini efektif pada data berdimensi tinggi seperti representasi teks.

Decision Tree membangun struktur pohon keputusan berdasarkan fitur yang paling informatif. Model ini mudah dipahami karena menghasilkan aturan yang eksplisit. Random Forest merupakan pengembangan dari Decision Tree yang menggabungkan banyak pohon untuk meningkatkan stabilitas dan akurasi. Perbandingan singkat algoritma tersebut dapat dilihat pada tabel berikut.

Algoritma	Karakteristik	Kelebihan	Keterbatasan
Naive Bayes	Probabilistik sederhana	Cepat dan efisien	Asumsi independensi fitur
Logistic Regression	Model linear	Mudah diinterpretasi	Terbatas pada pola linear
SVM	Margin maksimum	Akurasi tinggi	Sensitif terhadap parameter
Decision Tree	Struktur pohon	Transparan	Rentan overfitting
Random Forest	Ensemble pohon	Stabil dan akurat	Kurang interpretatif

Feature engineering menjadi tahap penting dalam pendekatan ini. Peneliti memilih fitur yang paling relevan untuk meningkatkan performa model. Seleksi fitur dapat dilakukan menggunakan metode statistik atau berbasis model. Pengurangan fitur membantu mengurangi dimensi dan meningkatkan efisiensi komputasi.

Masalah overfitting sering muncul ketika model terlalu menyesuaikan diri dengan data latih. Model mungkin mencapai akurasi tinggi pada data pelatihan, tetapi gagal pada data baru. Teknik seperti cross-validation dan regularisasi membantu mengurangi risiko ini. Generalisasi model menjadi indikator penting dalam penelitian empiris.

Pendekatan machine learning memiliki beberapa kelebihan. Model lebih adaptif dibandingkan metode berbasis aturan. Sistem dapat belajar dari data aktual dan menyesuaikan diri dengan variasi bahasa. Namun pendekatan ini membutuhkan data berlabel dalam jumlah cukup besar. Proses pelatihan juga memerlukan waktu dan sumber daya komputasi.

E. HYBRID APPROACH

Pendekatan hibrida muncul sebagai respons terhadap keterbatasan metode tunggal. Sistem ini menggabungkan kekuatan pendekatan leksikon dan machine learning. Integrasi dilakukan untuk meningkatkan akurasi sekaligus menjaga interpretabilitas. Pendekatan ini banyak digunakan ketika data berlabel terbatas tetapi sumber leksikal tersedia.

Konsep integrasi leksikon dan machine learning cukup sederhana. Kamus sentimen digunakan untuk menghasilkan fitur tambahan. Fitur tersebut kemudian dimasukkan ke dalam model klasifikasi statistik. Dengan cara ini, model tidak hanya belajar dari frekuensi kata, tetapi juga dari informasi polaritas yang telah ditentukan sebelumnya. Integrasi ini membantu model menangkap pola evaluatif secara lebih terarah.

Kombinasi aturan linguistik dan model statistik juga sering diterapkan. Aturan digunakan untuk menangani kasus khusus seperti negasi atau intensifier. Model statistik kemudian memproses fitur yang telah disesuaikan tersebut. Pendekatan ini menjaga fleksibilitas model sekaligus mempertahankan kontrol linguistik pada tahap awal. Strategi ini sering menghasilkan peningkatan performa pada data yang kompleks.

Strategi penggabungan dapat dilakukan dalam beberapa cara. Sistem dapat menjumlahkan skor leksikon dan skor probabilistik model. Sistem juga dapat memasukkan skor leksikon sebagai fitur numerik tambahan. Alternatif lain adalah menggunakan pendekatan dua tahap, di mana hasil rule-based menjadi input awal bagi model pembelajaran. Perbandingan strategi tersebut dapat dilihat pada tabel berikut.

Strategi	Mekanisme	Kelebihan	Keterbatasan
Skor Gabungan	Menjumlahkan skor leksikon dan model	Implementasi sederhana	Sensitif terhadap bobot
Fitur Tambahan	Skor leksikon menjadi fitur model	Fleksibel	Perlu penyesuaian parameter

Dua Tahap	Rule-based sebagai filter awal	Kontrol linguistik kuat	Kompleksitas meningkat

Beberapa studi menunjukkan bahwa pendekatan hibrida efektif pada domain dengan data terbatas. Dalam analisis ulasan produk lokal, kamus domain-spesifik membantu meningkatkan akurasi model statistik. Pada data media sosial dengan variasi bahasa tinggi, aturan sederhana untuk negasi dapat memperbaiki hasil klasifikasi. Pendekatan ini sering menghasilkan performa lebih stabil dibandingkan metode tunggal.

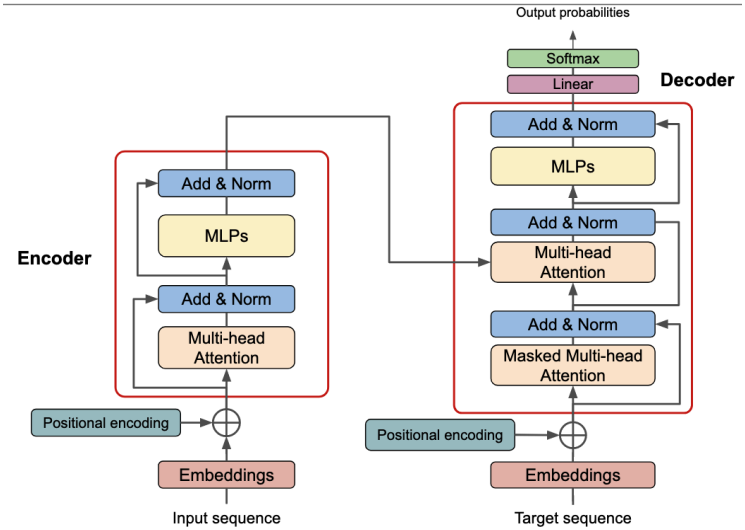
Keunggulan utama pendekatan hibrida terletak pada kemampuannya memanfaatkan sumber daya yang tersedia. Sistem tetap dapat bekerja meskipun dataset berlabel tidak terlalu besar. Integrasi leksikon membantu mengurangi ketergantungan penuh pada data latih. Pendekatan ini juga lebih mudah disesuaikan dengan domain tertentu.

Namun integrasi metode tidak selalu mudah. Penentuan bobot antara skor leksikon dan skor model membutuhkan eksperimen yang cermat. Konflik antara aturan linguistik dan prediksi model dapat muncul. Kompleksitas sistem juga meningkat karena melibatkan lebih dari satu mekanisme analisis. Tanpa perancangan yang hati-hati, integrasi justru dapat menurunkan konsistensi hasil.

Pendekatan hibrida menunjukkan bahwa tidak ada metode tunggal yang selalu optimal. Integrasi yang tepat dapat memperkuat sistem analisis sentimen, terutama dalam konteks penelitian terapan dengan keterbatasan data dan sumber daya.

F. DEEP LEARNING APPROACH

Pendekatan deep learning membawa perubahan mendasar dalam analisis sentimen. Berbeda dari machine learning klasik, model ini tidak bergantung pada rekayasa fitur manual secara intensif. Jaringan saraf dalam mampu mempelajari representasi fitur secara otomatis dari data mentah. Proses pembelajaran berlangsung melalui banyak lapisan transformasi nonlinier.



Ilustrasi Arsotektur Model LLM Deep Learning

(https://re-cinq.com/blog/llm-architectures)

Perbedaan utama terletak pada cara model menangkap pola. Machine learning klasik sering menggunakan fitur eksplisit seperti frekuensi kata. Deep learning menggunakan embedding yang memetakan kata ke dalam ruang vektor berdimensi rendah. Representasi ini memungkinkan model menangkap hubungan semantik antar kata secara lebih halus. Embedding juga membantu mengurangi masalah dimensi tinggi pada teks.

Model deep learning untuk teks berkembang dalam beberapa arsitektur utama. Convolutional Neural Network digunakan untuk menangkap pola lokal dalam urutan kata. Model ini efektif mengenali frasa penting dalam kalimat pendek. Recurrent Neural Network dirancang untuk memproses data berurutan dan mempertahankan informasi konteks sebelumnya. Model ini lebih sesuai untuk teks dengan dependensi panjang.

LSTM dan GRU merupakan pengembangan dari RNN yang mengatasi masalah vanishing gradient. Kedua model ini menggunakan mekanisme gerbang untuk mengatur aliran informasi. Dengan mekanisme tersebut, model dapat mempertahankan konteks dalam kalimat yang lebih panjang. Banyak penelitian menunjukkan bahwa LSTM memberikan hasil yang stabil dalam klasifikasi sentimen.

Transformer memperkenalkan mekanisme perhatian yang memproses seluruh urutan kata secara paralel. Model ini tidak lagi bergantung pada urutan

linear seperti RNN. Transformer mampu menangkap hubungan jarak jauh antar kata dengan lebih efisien. Arsitektur ini menjadi dasar bagi banyak model pralatih modern.

Ilustrasi Arsitektur Deep Learning

Fine-tuning dan transfer learning menjadi strategi penting dalam pendekatan modern. Model besar dilatih terlebih dahulu pada korpus umum berskala besar. Model tersebut kemudian disesuaikan pada dataset spesifik melalui proses fine-tuning. Pendekatan ini mengurangi kebutuhan data berlabel dalam jumlah sangat besar. Transfer learning juga mempercepat proses pengembangan model.

Meskipun menawarkan performa tinggi, pendekatan ini memiliki kebutuhan komputasi yang signifikan. Pelatihan model dalam membutuhkan perangkat keras dengan kemampuan pemrosesan paralel. Dataset besar sering diperlukan untuk mencapai performa optimal. Keterbatasan infrastruktur dapat menjadi hambatan bagi institusi kecil.

Interpretabilitas menjadi isu penting dalam deep learning. Model neural sering dianggap sebagai kotak hitam karena proses internalnya sulit dijelaskan secara langsung. Peneliti berupaya mengembangkan teknik visualisasi perhatian dan analisis bobot untuk meningkatkan transparansi. Namun interpretasi hasil tetap memerlukan kehati-hatian.

BAB 4. DATA TEKS DAN SUMBER INFORMASI UNTUK ANALISIS SENTIMEN

Deny Prasetyo, S.Pd., M.Kom.

Kualitas dan karakteristik data sangat memengaruhi hasil model. Pemilihan sumber data perlu disesuaikan dengan tujuan penelitian dan konteks domain. Setiap sumber memiliki pola bahasa, struktur, dan tingkat kebisingan yang berbeda.

A. KARAKTERISTIK UMUM DATA TEKS UNTUK ANALISIS SENTIMEN

Data teks yang digunakan dalam analisis sentimen umumnya termasuk kategori tidak terstruktur. Data terstruktur memiliki format tetap seperti tabel dengan kolom yang jelas. Data tidak terstruktur tidak memiliki skema baku dan sering berbentuk kalimat bebas. Komentar media sosial dan ulasan produk termasuk dalam kategori ini. Perbedaan keduanya dapat diringkas dalam tabel berikut.

Aspek	Data Terstruktur	Data Tidak Terstruktur
Format	Tabel atau skema tetap	Teks bebas
Kemudahan analisis	Lebih mudah diproses langsung	Perlu pra-pemrosesan
Contoh	Data transaksi	Komentar pengguna
Tantangan utama	Konsistensi format	Ambiguitas bahasa

Teks digital memiliki ciri khas yang membedakannya dari teks formal. Bahasa yang digunakan sering tidak baku dan singkat. Pengguna media sosial banyak menggunakan singkatan, emoji, dan campuran bahasa. Struktur kalimat sering tidak lengkap. Kondisi ini menuntut proses pembersihan dan normalisasi sebelum analisis dilakukan.

Variasi panjang dan gaya bahasa juga menjadi karakteristik penting. Ulasan produk dapat terdiri dari satu kalimat pendek atau paragraf panjang. Komentar berita sering memuat opini emosional dengan gaya retoris. Jawaban survei cenderung lebih formal dan terarah. Perbedaan ini memengaruhi strategi representasi dan pemodelan yang digunakan.

Noise dalam data media sosial menjadi tantangan utama. Noise dapat berupa kesalahan ejaan, pengulangan karakter, atau simbol yang tidak relevan. Spam dan komentar otomatis juga sering muncul dalam dataset besar. Tanpa penyaringan yang tepat, noise dapat menurunkan akurasi model. Proses pra-pemrosesan berperan penting untuk mengurangi gangguan ini.

Alur pengolahan data mentah hingga siap dianalisis dapat digambarkan sebagai berikut.

Alur Pengolahan Data

Kualitas data memiliki hubungan langsung dengan performa model. Dataset yang bersih dan konsisten membantu model mengenali pola secara lebih akurat. Dataset yang penuh noise dapat menghasilkan model yang bias atau tidak stabil. Selain itu, distribusi label yang seimbang juga memengaruhi hasil evaluasi. Oleh karena itu, tahap pengelolaan data tidak dapat dipandang sebagai langkah teknis semata.

B. MEDIA SOSIAL SEBAGAI SUMBER DATA

Media sosial menjadi salah satu sumber utama data dalam analisis sentimen. Platform seperti Twitter atau X, Instagram, dan Facebook menyediakan ruang bagi pengguna untuk menyampaikan opini secara terbuka. Setiap hari jutaan komentar dipublikasikan dalam berbagai topik. Volume dan kecepatan produksi teks membuat media sosial menarik bagi peneliti.

Twitter atau X dikenal dengan format pesan pendek yang padat. Pengguna sering menyampaikan opini dalam satu atau dua kalimat. Instagram lebih menekankan visual, tetapi bagian komentar tetap menjadi ruang diskusi publik. Facebook menyediakan status panjang dan kolom komentar yang beragam. Perbedaan karakter platform memengaruhi gaya bahasa yang digunakan.

Karakteristik utama teks media sosial adalah pendek dan interaktif. Banyak unggahan berupa tanggapan langsung terhadap peristiwa tertentu. Percakapan berlangsung dalam bentuk balasan berantai. Struktur dialog ini menciptakan konteks yang saling terhubung antar komentar. Analisis sentimen perlu mempertimbangkan konteks percakapan, bukan hanya teks tunggal.

Media Sosial Sebagai Sumber Data

Penggunaan hashtag, mention, emoji, dan singkatan menjadi ciri khas lain. Hashtag membantu mengelompokkan topik diskusi. Mention menghubungkan percakapan dengan akun tertentu. Emoji memperkuat ekspresi emosi tanpa kata tambahan. Singkatan dan bahasa gaul sering muncul untuk menghemat karakter. Elemen-elemen ini dapat membawa informasi sentimen yang penting. Peran unsur khas media sosial dapat diringkas dalam tabel berikut.

Unsur	Fungsi	Dampak pada Analisis
Hashtag	Penanda topik	Membantu klasifikasi tema
Mention	Referensi akun	Menunjukkan arah komunikasi
Emoji	Ekspresi emosi	Menambah konteks sentimen
Singkatan	Efisiensi bahasa	Menambah kompleksitas normalisasi

Media sosial juga mencerminkan dinamika opini publik secara real-time. Ketika suatu peristiwa terjadi, respons publik muncul dalam hitungan menit. Peneliti dapat memantau perubahan sentimen dari waktu ke waktu. Analisis temporal membantu memahami pola dukungan atau penolakan terhadap isu tertentu. Kecepatan ini menjadi keunggulan dibandingkan survei konvensional.

Namun terdapat tantangan serius terkait bias dan representasi pengguna. Tidak semua kelompok masyarakat aktif di media sosial. Algoritma platform

juga memengaruhi distribusi konten yang terlihat. Beberapa opini mungkin lebih menonjol karena efek viral. Dataset yang diambil tanpa pertimbangan representasi dapat menghasilkan kesimpulan yang bias.

Alur umum pemanfaatan data media sosial dapat digambarkan sebagai berikut.

Alur Pemanfaatan Data Media Sosial

Media sosial menyediakan peluang besar bagi penelitian analisis sentimen. Namun interpretasi hasil perlu dilakukan secara hati-hati. Data yang luas dan cepat tidak selalu mencerminkan keseluruhan populasi. Pemahaman terhadap karakter dan keterbatasan platform menjadi kunci dalam pemanfaatan data ini.

C. DATA SURVEI BERBASIS TEKS

Data survei berbasis teks berasal dari jawaban terbuka dalam kuesioner. Responden menuliskan pendapat dengan kata mereka sendiri.

SURVEY DATA COLLECTION METHODS	
Online Surveys	▪ Utilize digital platforms for convenience and quick data collection ▪ Suitable for various survey lengths and purposes
Face-to-face Surveys	▪ Involve direct interaction for in-depth exploration ▪ Ideal for complex topics but resource-intensive
Telephone Surveys	▪ Efficient for reaching dispersed audiences but may have limitations ▪ Immediate clarification of questions
Paper Surveys	▪ Physical questionnaires for traditional preferences ▪ Require manual data entry and analysis
Email Surveys	▪ Versatile for longer surveys, success depends on engagement ▪ Cost-effective with follow-up capabilities
Mobile Surveys	▪ Designed for mobile devices, ensuring widespread participation ▪ Capitalize on increased engagement and response rates
SMS Surveys	▪ Involve concise questions via text messages ▪ Effective for quick feedback, limited space for elaborate responses

Macam Macam Data Survey

Bentuk ini berbeda dari pertanyaan tertutup yang hanya memilih skala angka. Jawaban terbuka memberi ruang untuk alasan, kritik, dan saran yang lebih rinci.

Dalam konteks akademik, survei sering digunakan untuk mengevaluasi proses pembelajaran atau layanan kampus. Dalam kebijakan publik, survei

membantu pemerintah memahami respons masyarakat terhadap program tertentu. Data yang diperoleh biasanya terikat pada tujuan penelitian yang jelas. Peneliti juga mengetahui karakteristik responden seperti usia atau latar pendidikan. Informasi ini membantu interpretasi hasil secara lebih terarah.

Data survei memiliki perbedaan penting dibandingkan data media sosial. Media sosial bersifat spontan dan tidak terkontrol. Survei dirancang dengan instrumen yang sistematis dan pertanyaan yang spesifik. Media sosial mencerminkan opini publik yang luas tetapi tidak selalu terwakili secara proporsional. Survei memungkinkan pemilihan sampel yang lebih terstruktur.

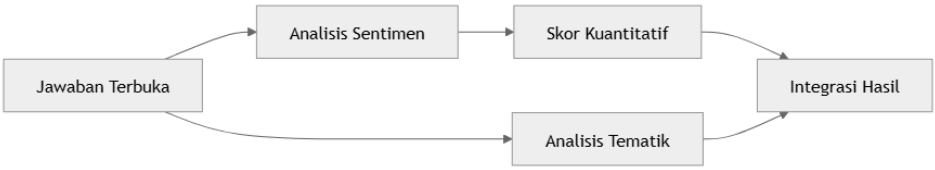
Perbandingan keduanya dapat diringkas dalam tabel berikut.

Aspek	Data Survei	Media Sosial
Struktur pertanyaan	Terencana dan terkontrol	Spontan
Karakter responden	Diketahui	Tidak selalu jelas
Representasi sampel	Dapat dirancang	Bergantung pada pengguna aktif
Volume data	Lebih terbatas	Sangat besar

Keunggulan utama data survei terletak pada kontrol variabel penelitian. Peneliti dapat menentukan populasi dan teknik sampling. Pertanyaan dapat dirancang untuk menggali aspek tertentu secara mendalam. Lingkungan yang lebih terkontrol membantu mengurangi bias eksternal. Hasil analisis sentimen dari survei sering lebih mudah dikaitkan dengan variabel demografis atau kebijakan tertentu.

Data survei juga membuka peluang integrasi analisis kuantitatif dan kualitatif. Skor sentimen dapat dihitung untuk setiap jawaban terbuka. Hasil tersebut kemudian dikaitkan dengan data numerik seperti nilai kepuasan atau tingkat partisipasi. Pendekatan ini memperkaya interpretasi karena menggabungkan angka dan narasi.

Alur integrasi analisis dapat digambarkan sebagai berikut.



Alur Analisis Hasil Survey

Pendekatan ini menunjukkan bahwa analisis sentimen tidak selalu berdiri sendiri. Dalam survei, metode ini dapat menjadi bagian dari strategi penelitian campuran. Dengan desain yang tepat, data survei berbasis teks memberikan informasi yang kaya dan terukur untuk mendukung pengambilan keputusan.

D. DATASET PUBLIK UNTUK PENELITIAN

Dataset publik memegang peran penting dalam pengembangan dan evaluasi model analisis sentimen. Dataset ini menyediakan data berlabel yang dapat digunakan secara luas oleh peneliti. Penggunaan dataset publik juga mendukung reproduksibilitas penelitian. Hasil yang diperoleh dapat dibandingkan secara objektif dengan studi lain.

IMDB Movie Review Dataset menjadi salah satu dataset klasik dalam klasifikasi sentimen. Dataset ini berisi ribuan ulasan film yang telah diberi label positif dan negatif. Teksnya relatif panjang dan kaya ekspresi opini. Dataset ini sering digunakan untuk menguji model klasifikasi biner. Struktur datanya cukup bersih sehingga cocok untuk eksperimen awal.

Amazon Product Review Dataset mencakup ulasan berbagai kategori produk. Dataset ini biasanya dilengkapi dengan rating numerik. Hubungan antara rating dan teks memungkinkan analisis korelasi antara skor dan ekspresi bahasa. Namun variasi domain dalam dataset ini cukup luas. Model yang dilatih pada satu kategori belum tentu optimal pada kategori lain.

Stanford Sentiment Treebank menawarkan pendekatan yang lebih rinci. Dataset ini tidak hanya memberi label pada dokumen, tetapi juga pada frasa dan kalimat. Struktur pohon sintaksis digunakan untuk menganalisis sentimen pada level yang lebih halus. Dataset ini sering digunakan untuk penelitian berbasis deep learning. Kompleksitas anotasinya membantu pengujian model pada level granular.

Untuk konteks bahasa Indonesia, IndoNLU menyediakan beberapa dataset yang relevan. Dataset ini mencakup berbagai tugas NLP termasuk analisis sentimen. Keberadaan dataset lokal penting karena struktur bahasa berbeda dari bahasa Inggris. Selain IndoNLU, terdapat dataset ulasan produk lokal dan komentar media sosial yang dikembangkan oleh peneliti dalam negeri. Ketersediaan data bahasa Indonesia masih terus berkembang.

Perbandingan beberapa dataset publik dapat dilihat pada tabel berikut.

Dataset	Bahasa	Unit Label	Karakteristik
IMDB	Inggris	Dokumen	Ulasan film panjang

Amazon	Inggris	Dokumen dan rating	Multidomain produk
Stanford Treebank	Inggris	Frasa dan kalimat	Label granular
IndoNLU	Indonesia	Dokumen	Dataset multi-tugas

Pemilihan dataset perlu mempertimbangkan beberapa kriteria. Bahasa menjadi faktor utama agar sesuai dengan konteks penelitian. Ukuran dataset memengaruhi stabilitas pelatihan model. Skema anotasi juga perlu diperhatikan karena setiap dataset memiliki definisi label yang berbeda. Selain itu, kesesuaian domain sangat menentukan relevansi hasil.

Dataset publik memiliki kelebihan yang jelas. Peneliti dapat menghemat waktu pengumpulan dan pelabelan data. Dataset ini juga memudahkan perbandingan performa model secara standar. Namun terdapat keterbatasan yang perlu dicermati. Beberapa dataset tidak merepresentasikan konteks lokal tertentu. Distribusi label mungkin tidak seimbang. Selain itu, data yang terlalu sering digunakan dapat menyebabkan model terlalu disesuaikan dengan pola tertentu.

Penggunaan dataset publik sebaiknya disertai pemahaman konteks dan batasannya. Dataset bukan hanya kumpulan teks berlabel, tetapi representasi situasi sosial tertentu. Pemilihan yang tepat akan meningkatkan relevansi dan validitas penelitian analisis sentimen.

E. TEKNIK PENGAMBILAN DATA

Pengambilan data menjadi tahap awal sebelum analisis sentimen dilakukan. Teknik yang digunakan harus mempertimbangkan legalitas dan kualitas data. Peneliti perlu memahami cara kerja platform digital dan format data yang dihasilkan. Proses ini tidak hanya bersifat teknis, tetapi juga metodologis.

Penggunaan API resmi platform merupakan metode yang paling disarankan. API menyediakan akses terstruktur terhadap data tertentu sesuai kebijakan penyedia layanan. Platform seperti Twitter atau X menyediakan endpoint untuk mengambil tweet berdasarkan kata kunci atau waktu tertentu. API biasanya membatasi jumlah permintaan dalam periode tertentu. Batasan ini perlu diperhitungkan dalam desain penelitian.

Web scraping menjadi alternatif ketika API tidak tersedia atau terbatas. Teknik ini mengambil data langsung dari halaman web dengan membaca struktur HTML. Peneliti menggunakan skrip untuk mengekstrak elemen teks tertentu. Metode ini memerlukan perhatian khusus terhadap kebijakan

penggunaan situs. Pengambilan data tanpa izin dapat melanggar ketentuan layanan.

Perbandingan penggunaan API dan web scraping dapat dilihat pada tabel berikut.

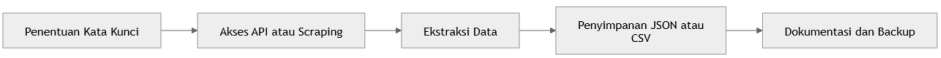
Aspek	API Resmi	Web Scraping
Legalitas	Lebih aman jika sesuai kebijakan	Perlu kehati-hatian
Struktur data	Terformat dan konsisten	Bergantung pada struktur halaman
Batas akses	Ada pembatasan kuota	Bergantung pada sistem situs
Kompleksitas teknis	Relatif terstandar	Perlu penyesuaian kode

Data yang diperoleh dari API umumnya berbentuk JSON. Format ini menyimpan data dalam pasangan kunci dan nilai. JSON memudahkan pemrosesan karena struktur hierarkisnya jelas. Format CSV juga sering digunakan untuk penyimpanan sederhana dalam bentuk tabel. Pemilihan format bergantung pada kebutuhan analisis dan kemudahan integrasi dengan perangkat lunak statistik.

Proses penyimpanan dan dokumentasi data memegang peran penting dalam penelitian. Data mentah sebaiknya disimpan terpisah dari data yang telah diproses. Dokumentasi perlu mencatat tanggal pengambilan, kata kunci, serta parameter pencarian. Catatan ini membantu menjaga transparansi dan reproduksibilitas. Tanpa dokumentasi yang jelas, penelitian sulit diverifikasi.

Otomatisasi pengumpulan data sering dilakukan untuk penelitian skala besar. Skrip dapat dijalankan secara berkala untuk mengambil data baru. Otomatisasi membantu memantau perubahan sentimen dari waktu ke waktu. Namun proses ini perlu diawasi agar tidak melampaui batas kebijakan platform.

Alur umum teknik pengambilan data dapat digambarkan sebagai berikut.



Alur Pengambulan Data

Teknik pengambilan data yang tepat akan meningkatkan kualitas analisis sentimen. Kesalahan pada tahap ini dapat berdampak pada bias atau ketidaklengkapan dataset. Oleh karena itu, perencanaan dan pencatatan yang cermat menjadi bagian penting dalam penelitian berbasis data teks.

F. ETIKA DAN LEGALITAS PENGAMBILAN DATA

Pengambilan data teks tidak dapat dilepaskan dari aspek etika dan hukum. Peneliti perlu memastikan bahwa proses pengumpulan dan penggunaan data tidak melanggar hak individu. Dalam konteks Indonesia, kerangka hukum utama mencakup Undang-Undang Informasi dan Transaksi Elektronik serta Undang-Undang Perlindungan Data Pribadi. Kedua regulasi ini mengatur pemanfaatan informasi digital dan perlindungan data pribadi.

Privasi dan perlindungan data pengguna menjadi prinsip utama. Data yang memuat nama, alamat surel, nomor telepon, atau identitas lain termasuk data pribadi. Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi mengatur bahwa pemrosesan data pribadi harus memiliki dasar hukum yang jelas. Peneliti perlu memastikan bahwa data yang digunakan tidak mengungkap identitas individu tanpa izin. Pelanggaran terhadap ketentuan ini dapat menimbulkan sanksi administratif maupun pidana.

Persetujuan menjadi aspek penting dalam pengumpulan data. Dalam penelitian survei, persetujuan responden biasanya diperoleh melalui lembar persetujuan tertulis. Pada data media sosial, situasinya lebih kompleks karena data bersifat publik tetapi tetap mengandung informasi pribadi. Praktik yang dianjurkan adalah melakukan anonimisasi sebelum analisis dilakukan. Proses ini menghapus atau menyamarkan identitas pengguna agar tidak dapat dilacak kembali.

Kebijakan platform dan terms of service juga harus diperhatikan. Setiap platform digital memiliki aturan mengenai penggunaan data dan akses API. Undang-Undang Nomor 19 Tahun 2016 tentang Perubahan atas Undang-Undang ITE menegaskan pentingnya kepatuhan terhadap sistem elektronik yang berlaku. Pengambilan data yang melanggar kebijakan platform dapat dianggap sebagai akses tanpa hak. Peneliti perlu membaca dan mematuhi ketentuan layanan sebelum melakukan scraping atau penggunaan API.

Risiko bias dan penyalahgunaan data juga perlu diantisipasi. Dataset yang diambil dari platform tertentu mungkin tidak mewakili seluruh populasi. Interpretasi hasil tanpa mempertimbangkan bias dapat menghasilkan kesimpulan yang keliru. Penyalahgunaan data untuk tujuan di luar penelitian juga dapat merugikan pihak tertentu. Oleh karena itu, analisis perlu dilakukan secara proporsional dan bertanggung jawab.

Prinsip transparansi dan reproducibility menjadi standar dalam penelitian ilmiah. Peneliti perlu menjelaskan sumber data, metode pengambilan, serta proses pembersihan yang dilakukan. Dokumentasi yang jelas membantu peneliti lain mereplikasi studi dengan kondisi serupa. Transparansi juga memperkuat akuntabilitas ketika penelitian berkaitan dengan isu publik yang sensitif.

Alur etis pengambilan data dapat digambarkan sebagai berikut.

Alur Legalitas Pengambilan Data

Etika dan legalitas bukan sekadar pelengkap metodologi. Aspek ini menentukan legitimasi dan kepercayaan terhadap hasil penelitian. Kepatuhan terhadap regulasi Indonesia membantu menjaga keseimbangan antara inovasi teknologi dan perlindungan hak individu.

BAB 5. PRA-PEMROSESAN DATA UNTUK ANALISIS SENTIMEN

Nurul Badriah, S.Kom., M.Kom.

A. PERAN PRA-PEMROSESAN

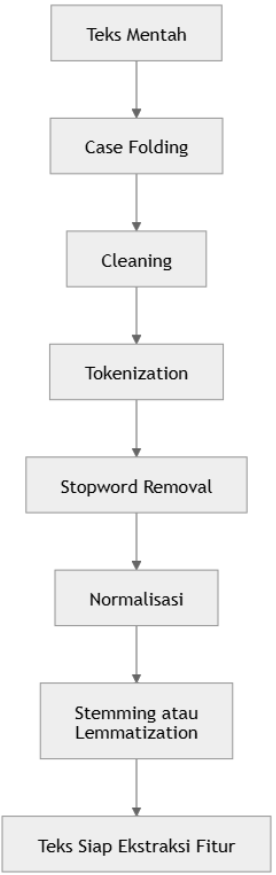
Pra-pemrosesan data teks merupakan tahap awal dalam sistem analisis sentimen. Tahap ini menyiapkan data mentah agar siap diproses oleh algoritma klasifikasi. Data teks yang berasal dari media sosial dan platform digital umumnya belum terstruktur. Sistem tidak dapat langsung mengolah data tersebut tanpa pembersihan dan normalisasi.

Pra-pemrosesan dapat didefinisikan sebagai serangkaian langkah untuk membersihkan, menstandarkan, dan mentransformasikan teks. Proses ini mengubah teks mentah menjadi bentuk yang lebih konsisten. Setiap kata disiapkan agar dapat direpresentasikan dalam bentuk numerik. Dengan demikian, model pembelajaran mesin dapat mengenali pola yang relevan.

Tujuan utama pra-pemrosesan dalam analisis sentimen adalah meningkatkan kualitas fitur. Proses ini mengurangi noise yang tidak memiliki makna evaluatif. Tahap ini juga membantu menurunkan dimensi fitur yang tidak penting. Selain itu, normalisasi kata mencegah duplikasi representasi akibat variasi ejaan. Semua langkah tersebut berkontribusi pada stabilitas model.

Kualitas data memiliki dampak langsung terhadap akurasi model. Dataset yang bersih cenderung menghasilkan performa lebih baik. Sebaliknya, data yang penuh kesalahan ejaan dan simbol acak dapat menurunkan presisi dan recall. Model dapat belajar dari pola yang keliru jika data tidak diproses dengan benar. Kondisi ini sering menyebabkan hasil klasifikasi yang tidak konsisten.

Data teks dalam analisis sentimen memiliki karakteristik tidak terstruktur. Teks tidak memiliki format baku seperti tabel numerik. Struktur kalimat sering bervariasi dan tidak selalu lengkap. Bahasa informal dan campuran bahasa sering muncul dalam satu kalimat. Ambiguitas makna juga menjadi tantangan tersendiri.



Alur Pra-Pemrosesan

Pipeline tersebut menunjukkan bahwa pra-pemrosesan bersifat bertahap dan sistematis. Setiap tahap memiliki fungsi spesifik dalam menyiapkan data. Proses ini menjadi fondasi sebelum representasi fitur dan pemodelan dilakukan. Tanpa fondasi yang kuat, sistem analisis sentimen sulit mencapai hasil yang akurat dan dapat dipercaya.

Sebagai contoh, kalimat “pelayanannya cepat sih, tapi bikin kesel” mengandung dua nuansa berbeda. Tanpa pemisahan dan analisis yang tepat, sistem dapat gagal menangkap dominasi sentimen negatif. Contoh lain muncul pada kalimat “produk ini BAGUS bangettt!!! 😍”. Jika huruf kapital, pengulangan huruf, dan emoji tidak diproses, model dapat memperlakukan variasi tersebut sebagai fitur berbeda. Akibatnya, distribusi kata menjadi tidak efisien.

Kegagalan pra-pemrosesan sering menghasilkan sparsity tinggi pada matriks fitur. Kata dengan makna sama dapat tercatat sebagai entitas berbeda. Dimensi fitur meningkat tanpa memberikan informasi tambahan. Model kemudian membutuhkan waktu pelatihan lebih lama dan menghasilkan performa yang kurang stabil.

Contoh Kalimat di Media Sosial yang Harus di Pra-Pemrosesan

B. KARAKTERISTIK DATA TEKS

Data teks dalam analisis sentimen memiliki sifat yang berbeda dari data numerik. Data ini tidak memiliki format baku seperti tabel atau angka terstruktur. Setiap dokumen dapat terdiri dari satu kalimat pendek atau beberapa

paragraf panjang. Struktur kalimat juga tidak selalu mengikuti aturan tata bahasa yang lengkap. Kondisi ini menuntut proses analisis yang lebih fleksibel.

Ambiguitas semantik menjadi salah satu tantangan utama. Satu kata dapat memiliki makna berbeda tergantung konteks. Kata "ringan" dapat berarti bobot yang kecil atau beban yang tidak berat secara psikologis. Tanpa pemahaman konteks, sistem dapat salah menafsirkan makna. Ambiguitas ini sering muncul dalam teks ulasan dan komentar daring.

Teks juga sangat sensitif terhadap konteks. Makna sebuah kalimat dapat berubah karena keberadaan kata negasi atau konjungsi. Kalimat "produknya bagus, tapi mahal" mengandung evaluasi campuran. Jika sistem hanya mendeteksi kata "bagus," hasilnya akan bias ke arah positif. Analisis sentimen perlu mempertimbangkan hubungan antar kata dalam satu struktur kalimat.

Variasi morfologi dan sintaksis memperumit proses analisis. Bahasa Indonesia memiliki banyak bentuk imbuhan seperti me-, di-, pe-, dan ke-. Kata "membeli," "dibeli," dan "pembelian" berasal dari akar yang sama. Tanpa proses normalisasi, variasi ini akan dianggap sebagai fitur berbeda. Struktur sintaksis juga dapat memengaruhi interpretasi sentimen.

Unsur emosional sering muncul secara implisit. Tidak semua opini disampaikan secara langsung. Kalimat seperti "lumayan lah" dapat bermakna netral atau sedikit negatif tergantung konteks. Sistem yang hanya mengandalkan kata positif dan negatif mungkin gagal menangkap nuansa tersebut. Interpretasi emosional memerlukan analisis yang lebih mendalam.

Teks digital memiliki elemen khas yang jarang ditemukan pada teks formal. Elemen ini memengaruhi cara sistem memproses data. Beberapa elemen utama dapat diringkas dalam tabel berikut.

Elemen	Fungsi dalam Teks	Dampak pada Analisis
URL	Referensi tautan	Umumnya dihapus
Hashtag	Penanda topik	Dapat mengandung sentimen
Mention	Referensi akun	Biasanya dihapus atau dinetralkan
Emoji	Ekspresi emosi	Sumber sinyal sentimen
Simbol khusus	Penekanan ekspresi	Perlu pembersihan
Kata slang	Bahasa informal	Perlu normalisasi

URL sering tidak memiliki nilai sentimen langsung dan biasanya dihapus. Hashtag dapat memuat opini atau topik tertentu. Mention lebih sering berfungsi

sebagai alamat akun dan tidak selalu relevan dengan sentimen. Emoji sering menjadi indikator emosi yang kuat dan perlu dipertimbangkan secara hati-hati.

Kata slang dan singkatan banyak muncul dalam media sosial. Bentuk seperti “gk,” “bgt,” atau “mantul” memerlukan normalisasi. Tanpa proses ini, sistem akan memperlakukan setiap variasi sebagai kata berbeda. Akibatnya, dimensi fitur meningkat tanpa peningkatan informasi.

Karakteristik tersebut menunjukkan bahwa data teks memiliki kompleksitas tinggi. Sistem analisis sentimen perlu dirancang dengan mempertimbangkan dinamika bahasa digital. Pra-pemrosesan yang tepat membantu mengurangi gangguan sekaligus mempertahankan makna evaluatif yang penting.

C. TAHAPAN DASAR PRA-PEMROSESAN

Tahapan dasar pra-pemrosesan bertujuan menyederhanakan bentuk teks tanpa menghilangkan makna utama. Proses ini dilakukan sebelum tokenisasi dan representasi fitur. Dua langkah awal yang paling umum adalah case folding dan cleaning. Keduanya membantu mengurangi variasi yang tidak perlu dalam data.

1. Case Folding

Case folding adalah proses mengubah seluruh huruf dalam teks menjadi huruf kecil. Proses ini bertujuan menghindari perbedaan fitur akibat kapitalisasi. Sistem komputer membedakan huruf besar dan kecil sebagai entitas berbeda. Tanpa case folding, kata “Bagus” dan “bagus” akan tercatat sebagai dua token berbeda.

Kapitalisasi sering muncul dalam teks digital sebagai bentuk penekanan. Pengguna dapat menulis “BAGUS” untuk menunjukkan antusiasme. Namun secara semantik, kata tersebut tetap merujuk pada makna yang sama. Jika kapitalisasi tidak diseragamkan, matriks fitur akan memiliki dimensi lebih besar tanpa penambahan informasi berarti.

Dampak kapitalisasi terhadap fitur dapat dilihat pada tabel berikut.

Teks Asli	Token Tanpa Case Folding	Token Dengan Case Folding
Produk Ini BAGUS	Produk, Ini, BAGUS	produk, ini, bagus
Pelayanan Bagus	Pelayanan, Bagus	pelayanan, bagus

Contoh transformasi sederhana dapat dilihat pada kalimat berikut.

Sebelum case folding:
“Produk Ini BAGUS Sekali”
Setelah case folding:
“produk ini bagus sekali”

Proses ini sederhana tetapi berdampak besar pada konsistensi data. Dimensi fitur menjadi lebih terkendali dan distribusi kata lebih stabil.

2. Cleaning

Cleaning adalah proses menghapus elemen yang tidak relevan dari teks. Data digital sering mengandung URL, mention, hashtag, angka, dan simbol khusus. Elemen tersebut tidak selalu berkontribusi pada analisis sentimen. Jika tidak dibersihkan, sistem dapat mengenali pola yang tidak bermakna.

Penghapusan URL menjadi langkah umum dalam cleaning. Tautan seperti “http://promo.com” biasanya tidak mengandung opini. Mention seperti “@username” juga sering tidak memiliki nilai sentimen langsung. Hashtag dapat dihapus atau dipertahankan tergantung konteks penelitian. Jika hashtag mengandung kata evaluatif, peneliti dapat memisahkan simbol pagar dan mempertahankan katanya.

Angka dan simbol khusus sering dihapus untuk mengurangi noise. Tanda baca berlebihan seperti “!!!” atau “???” juga biasanya dinormalisasi. Pengulangan simbol dapat memperbesar dimensi fitur tanpa memberikan informasi tambahan. Cleaning membantu menyederhanakan struktur teks sebelum proses berikutnya.

Contoh transformasi cleaning dapat dilihat berikut ini.
Teks asli:
“Produk ini bagus banget!!! 😍😍 http://promo.com”

Setelah cleaning dasar:
“produk ini bagus banget 😍😍”

Pada contoh tersebut, URL dan tanda seru berlebihan dihapus. Emoji tetap dipertahankan karena dapat menjadi indikator emosi. Dalam beberapa penelitian, emoji dikonversi menjadi token khusus seperti “emoji_positif.” Keputusan mempertahankan atau menghapus emoji perlu disesuaikan dengan tujuan analisis.

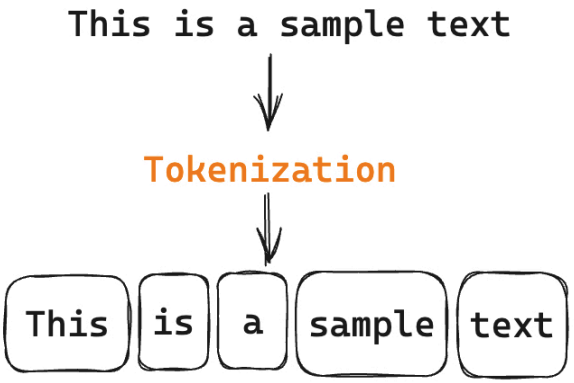
Case folding dan cleaning membentuk fondasi awal dalam pra-pemrosesan. Kedua tahap ini membantu menyederhanakan teks tanpa menghilangkan makna evaluatif. Dengan data yang lebih konsisten, tahap tokenisasi dan analisis berikutnya dapat dilakukan secara lebih efektif.

D. TOKENIZATION DAN STOPWORD REMOVAL

Setelah proses cleaning selesai, teks perlu dipecah menjadi unit yang lebih kecil. Unit ini menjadi dasar dalam pembentukan fitur numerik. Dua tahap penting pada bagian ini adalah tokenization dan stopword removal. Keduanya memengaruhi struktur representasi akhir dalam model.

1. Tokenization

Tokenization adalah proses memecah teks menjadi token. Token biasanya berupa kata, tetapi dapat juga berupa frasa atau karakter. Proses ini membantu sistem mengenali elemen dasar dalam dokumen. Tanpa tokenisasi, teks akan diperlakukan sebagai satu rangkaian karakter panjang.



Ilustrasi Tokenization

Fungsi utama tokenisasi adalah mengidentifikasi unit makna yang dapat dihitung. Setiap token nantinya menjadi kandidat fitur dalam matriks representasi. Tokenisasi juga memudahkan proses pencocokan dengan kamus sentimen. Selain itu, tahap ini membantu mengurangi kompleksitas struktur kalimat.

Tokenisasi berbasis spasi merupakan metode paling sederhana. Sistem memecah teks berdasarkan tanda spasi. Metode ini cukup efektif untuk teks formal yang rapi. Namun pendekatan ini memiliki keterbatasan ketika menghadapi tanda baca atau simbol khusus.

Tokenisasi berbasis aturan menggunakan pola tertentu untuk menentukan batas token. Sistem dapat menghapus tanda baca atau memisahkan kata yang diikuti simbol. Pendekatan ini lebih fleksibel dibandingkan metode berbasis spasi. Beberapa pustaka NLP menyediakan tokenizer yang mempertimbangkan bahasa dan struktur tertentu.

Contoh sederhana tokenisasi dapat dilihat berikut ini.

Teks:

“layanan sangat memuaskan”

Hasil tokenisasi:

[layanan, sangat, memuaskan]

Tantangan muncul pada bahasa informal. Kalimat seperti “bagusss bangettt” memuat pengulangan huruf. Tokenisasi sederhana akan tetap menghasilkan token yang belum dinormalisasi. Campuran bahasa seperti “produknya very good” juga memerlukan penanganan tambahan. Tanpa normalisasi, variasi ini akan memperbesar dimensi fitur.

2. Stopword Removal

Stopword adalah kata umum yang sering muncul dalam teks tetapi memiliki sedikit nilai informatif. Kata seperti “dan,” “yang,” “di,” dan “ke” termasuk kategori ini. Kata tersebut sering muncul dalam hampir semua dokumen. Jika tidak dihapus, stopword dapat mendominasi matriks fitur.

Daftar stopword biasanya disusun berdasarkan frekuensi kemunculan. Setiap bahasa memiliki daftar stopword yang berbeda. Dalam bahasa Indonesia, kata seperti “ini,” “itu,” dan “dari” sering dimasukkan dalam daftar. Penghapusan stopword membantu mengurangi dimensi fitur dan meningkatkan efisiensi komputasi.

Namun proses ini tidak selalu sederhana. Risiko terbesar adalah penghapusan kata negasi seperti “tidak” atau “bukan.” Kata tersebut memiliki pengaruh besar terhadap polaritas. Jika kata negasi dihapus, makna kalimat dapat berubah total. Kalimat “produk ini tidak bagus” akan berubah menjadi “produk ini bagus” setelah penghapusan negasi.

Perbandingan dampak penghapusan negasi dapat dilihat pada tabel berikut.

Kalimat Asli	Setelah Stopword Removal Umum	Dampak

Produk ini tidak bagus	produk bagus	Makna berubah
Layanan sangat cepat	layanan cepat	Makna tetap

Strategi selektif penghapusan menjadi solusi yang lebih aman. Peneliti dapat menyusun daftar stopword khusus untuk analisis sentimen. Kata negasi dan intensifier sebaiknya dipertahankan. Proses ini memerlukan evaluasi terhadap karakteristik dataset.

Tokenization dan stopword removal membentuk tahap penting dalam reduksi fitur. Proses ini membantu sistem fokus pada kata yang relevan dengan sentimen. Dengan pendekatan selektif, makna evaluatif tetap terjaga dan model dapat bekerja secara lebih akurat.

E. NORMALISASI, STEMMING, DAN LEMMATIZATION

Tahap ini bertujuan menyederhanakan variasi bentuk kata tanpa menghilangkan makna inti. Bahasa digital sering memuat bentuk tidak baku dan variasi ejaan. Jika variasi tersebut tidak ditangani, sistem akan menganggapnya sebagai fitur berbeda. Akibatnya, dimensi matriks fitur meningkat dan distribusi kata menjadi jarang.

1. Normalisasi Kata

Normalisasi teks adalah proses mengubah kata tidak baku menjadi bentuk standar. Proses ini penting pada data media sosial yang penuh singkatan dan variasi ejaan. Normalisasi membantu menyatukan bentuk kata yang memiliki makna sama. Dengan demikian, sistem dapat mengenali pola secara lebih konsisten.

Singkatan fonetik sering muncul dalam komunikasi digital. Contoh umum adalah “gk” untuk “tidak” dan “bgt” untuk “banget.” Tanpa normalisasi, kata tersebut akan dianggap entitas berbeda dari bentuk bakunya. Pengulangan huruf juga sering digunakan untuk penekanan emosi. Kata “baguuuusss” seharusnya dinormalisasi menjadi “bagus.”

Code-mixing menjadi tantangan lain dalam teks digital. Kalimat seperti “produk ini worth it banget” memadukan bahasa Inggris dan Indonesia. Sistem perlu menentukan apakah akan mempertahankan istilah asing atau menerjemahkannya. Typo juga sering muncul akibat kesalahan pengetikan. Kata “memuaskna” perlu dikoreksi menjadi “memuaskan.” Beberapa pendekatan

digunakan dalam normalisasi teks antaralain berbasis Kamus, Edit Distance, Character-level modelling, Seq2Seq, dan Transformer Based.

Pendekatan berbasis kamus menggunakan daftar pasangan kata tidak baku dan kata baku. Metode ini mudah diterapkan dan tidak membutuhkan pelatihan model. Namun kamus perlu diperbarui secara berkala karena bahasa digital terus berkembang.

Edit Distance atau Levenshtein Distance menghitung jarak antar kata berdasarkan jumlah operasi perubahan huruf. Jika jarak kecil, sistem menganggap kata tersebut variasi dari kata baku tertentu. Metode ini sederhana dan cepat untuk dataset kecil. Namun metode ini tidak mempertimbangkan konteks kalimat.

Character-level modeling bekerja pada tingkat karakter. Model mempelajari pola perubahan huruf dalam berbagai variasi ejaan. Pendekatan ini lebih tahan terhadap typo dan kata baru. Namun model membutuhkan data pelatihan yang cukup besar.

Sequence-to-Sequence atau Seq2Seq mengubah satu urutan kata menjadi urutan lain. Model ini banyak digunakan dalam penerjemahan mesin. Dalam normalisasi, model mengubah kalimat tidak baku menjadi kalimat baku. Metode ini fleksibel tetapi membutuhkan dataset paralel.

Transformer-based correction menggunakan model berbasis perhatian yang memahami konteks kalimat secara menyeluruh. Model ini mampu menangani ambiguitas dan slang kompleks. Namun kebutuhan komputasinya tinggi dan memerlukan proses fine-tuning.

Perbandingan pendekatan normalisasi dapat dilihat pada tabel berikut.

Metode	Konteks	Kompleksitas	Kebutuhan Data	Kelebihan	Keterbatasan
Kamus	Tidak	Rendah	Tidak perlu	Mudah diterapkan	Kurang fleksibel
Edit Distance	Tidak	Rendah	Tidak perlu	Cepat	Abaikan konteks
Character-Level	Terbatas	Sedang	Perlu	Tahan typo	Butuh data cukup
Seq2Seq	Ya	Tinggi	Perlu paralel	Fleksibel	Komputasi besar

Transformer	Sangat kontekstual	Sangat tinggi	Perlu fine-tuning	Akurasi tinggi	Mahal secara komputasi

Setiap metode memiliki kelebihan dan kekurangan. Pemilihan pendekatan bergantung pada skala data dan sumber daya yang tersedia.

2. Stemming

Stemming adalah proses mengurangi kata berimbuhan menjadi bentuk dasar. Proses ini tidak selalu mempertimbangkan konteks gramatikal. Tujuannya adalah menyatukan variasi morfologis agar fitur lebih ringkas.

Dalam bahasa Indonesia, algoritma Nazief-Adriani sering digunakan. Algoritma ini menghapus imbuhan berdasarkan aturan morfologi. Contoh perubahan kata dapat dilihat berikut ini.

Kata Asli	Hasil Stemming
membeli	beli
dibeli	beli
pembelian	beli
ceritanya	cerita

Stemming membantu mengurangi variasi kata turunan. Namun hasilnya tidak selalu sesuai bentuk kamus.

3. Lemmatization

Lemmatization berbeda dari stemming karena mempertimbangkan konteks gramatikal. Proses ini mengembalikan kata ke bentuk lema yang sesuai dalam kamus. Lemmatization biasanya menggunakan analisis morfologi dan informasi tata bahasa.

Pendekatan berbasis konteks memungkinkan sistem menentukan bentuk dasar secara lebih akurat. Sebagai contoh, kata “berlari” akan diubah menjadi “lari.” Jika kata memiliki bentuk berbeda dalam konteks tertentu, sistem dapat memilih bentuk yang tepat.

Keunggulan lemmatization terletak pada akurasi linguistik. Hasilnya lebih konsisten dibandingkan stemming. Namun proses ini lebih kompleks dan membutuhkan sumber daya tambahan. Dalam praktik analisis sentimen, pemilihan antara stemming dan lemmatization perlu disesuaikan dengan kebutuhan penelitian.

Normalisasi, stemming, dan lemmatization membantu menyederhanakan struktur teks. Ketiga tahap ini mengurangi variasi kata dan meningkatkan konsistensi fitur. Dengan teks yang lebih terstandar, tahap representasi numerik dapat dilakukan secara lebih efisien dan akurat.

F. TRANSFORMASI TEKS KE REPRESENTASI NUMERIK

Setelah teks melalui tahap pra-pemrosesan, langkah berikutnya adalah mengubahnya ke bentuk numerik. Model pembelajaran mesin tidak dapat memproses teks dalam bentuk string. Sistem membutuhkan representasi vektor agar perhitungan matematis dapat dilakukan. Tahap ini menentukan bagaimana informasi bahasa diterjemahkan menjadi angka.

Bag of Words merupakan metode paling dasar dalam representasi teks. Metode ini menghitung frekuensi kemunculan setiap kata dalam dokumen. Urutan kata diabaikan sehingga struktur kalimat tidak dipertimbangkan. Setiap dokumen direpresentasikan sebagai vektor berdasarkan jumlah kata unik dalam korpus. Pendekatan ini sederhana dan mudah diimplementasikan.

TF-IDF mengembangkan konsep Bag of Words dengan menambahkan bobot. Bobot dihitung berdasarkan frekuensi kata dalam dokumen dan kelangkaannya dalam seluruh korpus. Kata yang sering muncul di satu dokumen tetapi jarang muncul di dokumen lain akan memiliki nilai lebih tinggi. Pendekatan ini membantu menyoroti kata yang lebih informatif.

Rumus umum TF-IDF dapat dituliskan sebagai berikut.

Term Frequency (TF):

$$\mathrm{TF}(t, \text{text}) = \frac{\text{Number of times term } t \text{ appears in text}}{\text{Total number of terms in the text}}$$

Inverse Document Frequency (IDF):

$$\mathrm{IDF}(t, \text{corpus}) = \log\left(\frac{\text{Total number of texts in corpus } |\text{corpus}|}{\text{Number of texts containing term } t}\right)$$

TF-IDF:

$$\text{TF-IDF}(t, \text{text}, \text{corpus}) = \mathrm{TF}(t, \text{text}) \times \mathrm{IDF}(t, \text{corpus})$$

Metode ini banyak digunakan dalam model machine learning klasik.

Word embedding memperkenalkan representasi berbasis vektor kontinu. Setiap kata dipetakan ke dalam ruang berdimensi rendah. Representasi ini memungkinkan model menangkap hubungan semantik antar kata. Kata dengan makna serupa akan memiliki jarak yang dekat dalam ruang vektor. Pendekatan ini lebih mampu menangkap makna dibandingkan metode berbasis frekuensi.

Representasi berbasis transformer menggunakan embedding kontekstual. Setiap kata memiliki representasi berbeda tergantung pada kalimatnya. Model seperti ini mampu memahami konteks secara lebih mendalam. Hubungan antar kata dianalisis menggunakan mekanisme perhatian. Pendekatan ini sering memberikan performa tinggi pada tugas analisis sentimen.

Perbandingan metode representasi dapat dilihat pada tabel berikut.

Metode	Memperhatikan Urutan	Konteks	Dimensi	Kompleksitas
Bag of Words	Tidak	Tidak	Tinggi	Rendah
TF-IDF	Tidak	Tidak	Tinggi	Rendah
Word Embedding	Terbatas	Sebagian	Rendah	Sedang
Transformer	Ya	Kontekstual penuh	Rendah hingga sedang	Tinggi

Hubungan antara pra-pemrosesan dan sparsity fitur sangat erat. Jika teks tidak dinormalisasi, variasi ejaan akan memperbesar jumlah kata unik. Kondisi ini meningkatkan sparsity pada matriks fitur. Sparsity tinggi membuat model lebih sulit menemukan pola yang stabil. Pra-pemrosesan membantu menyatukan variasi kata sehingga representasi menjadi lebih padat.

Sebagai contoh, kata "bagus," "BAGUS," dan "baguuus" akan menjadi tiga fitur berbeda tanpa normalisasi. Setelah pra-pemrosesan, ketiganya dapat disatukan menjadi satu fitur. Dimensi matriks berkurang dan distribusi frekuensi menjadi lebih representatif. Dengan demikian, kualitas representasi numerik sangat bergantung pada kualitas tahap pra-pemrosesan sebelumnya.

G. TANTANGAN DALAM PRA-PEMROSESAN

Pra-pemrosesan tidak selalu berjalan sederhana. Teks digital mengandung banyak variasi dan pola yang sulit ditangani secara otomatis. Beberapa tantangan dapat memengaruhi kualitas hasil analisis sentimen. Tantangan ini perlu dipahami sebelum model dikembangkan.

Sarkasme menjadi salah satu masalah paling kompleks. Kalimat sarkastik sering menyatakan sesuatu yang berlawanan dengan makna literalnya. Contoh seperti "pelayanannya cepat banget, sampai dua jam nunggu" tampak positif pada awal kalimat. Namun makna keseluruhan bersifat negatif. Sistem berbasis kata kunci sering gagal mengenali pola ini.

Ambiguitas kata juga menimbulkan kesulitan. Satu kata dapat memiliki makna berbeda tergantung konteks. Kata "dingin" dapat berarti suhu rendah atau sikap tidak ramah. Tanpa analisis konteks yang memadai, sistem dapat salah menentukan polaritas. Ambiguitas ini sering muncul pada ulasan produk makanan atau layanan.

Code-mixing sering ditemukan dalam teks media sosial. Pengguna mencampur bahasa Indonesia dan Inggris dalam satu kalimat. Contoh seperti "produknya bagus sih, tapi packaging-nya very bad" memuat dua bahasa sekaligus. Jika sistem hanya dilatih pada satu bahasa, sebagian sentimen dapat terlewat. Normalisasi dan kamus multibahasa menjadi penting dalam situasi ini.

Data tidak seimbang juga menjadi tantangan serius. Dalam banyak dataset, jumlah komentar positif jauh lebih banyak daripada komentar negatif. Model cenderung belajar memprediksi kelas dominan. Akibatnya, komentar negatif dapat salah diklasifikasikan sebagai positif. Ketidakseimbangan ini memengaruhi evaluasi dan interpretasi hasil.

Spam mengganggu kualitas dataset. Kalimat seperti "klik link ini sekarang juga" tidak mengandung opini terhadap objek tertentu. Jika spam tidak

disaring, model dapat belajar pola yang tidak relevan. Proses cleaning perlu dirancang untuk mendeteksi dan menghapus konten semacam ini.

Duplikasi data juga dapat memengaruhi distribusi fitur. Komentar yang sama dapat muncul berkali-kali akibat bot atau promosi. Pengulangan ini meningkatkan bobot kata tertentu secara tidak proporsional. Model dapat menganggap kata tersebut sangat penting meskipun sebenarnya hanya hasil pengulangan.

Ringkasan tantangan dalam pra-pemrosesan dapat dilihat pada tabel berikut.

Tantangan	Dampak pada Analisis
Sarkasme	Salah interpretasi polaritas
Ambiguitas kata	Kesalahan klasifikasi konteks
Code-mixing	Deteksi sentimen tidak lengkap
Data tidak seimbang	Bias pada kelas dominan
Spam	Fitur tidak relevan
Duplikasi	Distorsi distribusi kata

Tantangan tersebut menunjukkan bahwa pra-pemrosesan bukan sekadar pembersihan teknis. Proses ini memerlukan pemahaman linguistik dan konteks sosial. Tanpa strategi yang tepat, kualitas model analisis sentimen dapat menurun meskipun algoritma yang digunakan sudah canggih.

BAB 6. REPRESENTASI FITUR DAN PEMBOBOTAN TEKS

Ardy Wicaksono, S.Kom., M.Kom.

A. KONSEP DASAR REPRESENTASI FITUR TEKS

Model komputasi tidak dapat memproses teks dalam bentuk kata atau kalimat mentah. Algoritma klasifikasi bekerja pada angka dan operasi matematis. Oleh karena itu, teks perlu diubah menjadi representasi numerik. Proses ini disebut sebagai ekstraksi atau representasi fitur.

Representasi numerik memungkinkan sistem menghitung jarak, kemiripan, dan probabilitas antar dokumen. Tanpa konversi ini, model tidak dapat mengenali pola sentimen. Kata seperti “bagus” atau “buruk” harus direpresentasikan sebagai nilai tertentu dalam vektor. Setiap dokumen akhirnya menjadi kumpulan angka yang mencerminkan distribusi kata.

Hubungan antara fitur dan algoritma klasifikasi sangat erat. Algoritma seperti Naive Bayes dan Logistic Regression bergantung pada frekuensi atau bobot kata. Support Vector Machine memanfaatkan struktur vektor untuk mencari batas pemisah antar kelas. Pada deep learning, embedding digunakan sebagai input bagi jaringan saraf. Pilihan representasi dapat memengaruhi hasil klasifikasi secara langsung.

Dimensi fitur menjadi aspek penting dalam desain sistem. Jika korpus memiliki sepuluh ribu kata unik, maka representasi berbasis frekuensi dapat menghasilkan sepuluh ribu dimensi. Dimensi tinggi meningkatkan kebutuhan memori dan waktu komputasi. Model juga berisiko mengalami overfitting ketika jumlah fitur terlalu besar dibandingkan jumlah data.

Distribusi kata dalam korpus sering tidak merata. Sebagian kecil kata muncul sangat sering, sementara sebagian besar jarang muncul. Pola ini dikenal sebagai distribusi Zipf. Akibatnya, matriks fitur biasanya mengandung banyak nilai nol. Kondisi ini disebut sparsity.

Sparsity memengaruhi efisiensi dan stabilitas model. Matriks yang sangat jarang dapat memperlambat proses pelatihan. Namun dalam beberapa algoritma

linear, sparsity masih dapat ditangani secara efisien. Tantangan muncul ketika dimensi meningkat tanpa penambahan informasi yang bermakna.

Ilustrasi sederhana matriks dokumen-term dapat dilihat berikut ini.

Dokumen	bagus	buruk	cepat	mahal
D1	2	0	1	0
D2	0	1	0	2
D3	1	0	0	1

Pada tabel tersebut, setiap baris mewakili dokumen dan setiap kolom mewakili kata. Nilai dalam sel menunjukkan frekuensi kemunculan kata. Banyak sel bernilai nol karena tidak semua kata muncul pada setiap dokumen. Kondisi ini mencerminkan sparsity yang umum terjadi dalam analisis teks.

B. BAG OF WORDS (BOW)

Bag of Words merupakan metode representasi teks yang paling sederhana dan banyak digunakan pada tahap awal penelitian. Prinsip dasarnya adalah merepresentasikan dokumen sebagai kumpulan kata tanpa memperhatikan urutan kemunculannya. Setiap kata dihitung berdasarkan frekuensi dalam dokumen. Model kemudian menggunakan frekuensi tersebut sebagai fitur numerik.

Pendekatan ini bekerja dengan membangun kosakata dari seluruh korpus. Setiap kata unik menjadi satu dimensi dalam vektor. Jika korpus memiliki seribu kata unik, maka setiap dokumen direpresentasikan sebagai vektor berdimensi seribu. Nilai pada setiap dimensi menunjukkan jumlah kemunculan kata tersebut.

Representasi berbasis frekuensi kata bersifat langsung dan mudah dipahami. Kata yang sering muncul dalam dokumen akan memiliki nilai lebih besar. Kata yang tidak muncul akan bernilai nol. Struktur ini menghasilkan matriks dokumen-term yang cenderung jarang karena sebagian besar kata tidak muncul di setiap dokumen.

BoW secara eksplisit mengabaikan urutan kata. Kalimat “produk ini tidak bagus” dan “produk ini bagus tidak” akan memiliki representasi yang sama jika hanya menghitung frekuensi kata. Pengabaian urutan membuat metode ini sederhana, tetapi juga membatasi kemampuannya dalam menangkap konteks. Informasi sintaksis dan relasi antar kata tidak dipertimbangkan.

Kelebihan BoW terletak pada kesederhanaan dan efisiensinya. Metode ini mudah diimplementasikan dan tidak membutuhkan pelatihan model kompleks.

BoW juga cocok untuk algoritma machine learning klasik seperti Naive Bayes dan Support Vector Machine. Pada dataset berukuran sedang, metode ini sering memberikan hasil yang kompetitif.

Namun terdapat beberapa keterbatasan. Dimensi fitur dapat menjadi sangat besar ketika kosakata bertambah. Sparsity meningkat seiring bertambahnya jumlah kata unik. Metode ini juga tidak mampu menangkap makna semantik atau hubungan antar kata. Kata sinonim seperti “baik” dan “bagus” diperlakukan sebagai entitas terpisah.

Contoh pembentukan matriks dokumen-term dapat dilihat berikut ini.

Misalkan terdapat tiga dokumen:

D1: “produk ini bagus dan cepat”
D2: “produk ini mahal dan buruk”
D3: “layanan cepat dan bagus”

Langkah pertama adalah membangun kosakata unik dari seluruh dokumen:
[produk, ini, bagus, dan, cepat, mahal, buruk, layanan]

Matriks dokumen-term yang dihasilkan dapat ditampilkan sebagai berikut.

Dokumen	**produk**	**ini**	**bagus**	**dan**	**cepat**	**mahal**	**buruk**	**layanan**
D1	1	1	1	1	1	0	0	0
D2	1	1	0	1	0	1	1	0
D3	0	0	1	1	1	0	0	1

Tabel tersebut menunjukkan bahwa setiap dokumen direpresentasikan sebagai vektor frekuensi. Banyak nilai nol muncul karena kata tertentu tidak hadir pada semua dokumen. Kondisi ini mencerminkan sifat sparse dari representasi BoW.

Bag of Words menjadi dasar bagi metode pembobotan lanjutan seperti TF dan TF-IDF. Meskipun sederhana, pendekatan ini tetap relevan dalam banyak studi analisis sentimen, terutama ketika interpretabilitas dan efisiensi menjadi prioritas utama.

C. TERM FREQUENCY (TF)

Term Frequency merupakan metode pembobotan kata yang paling dasar dalam representasi teks. TF mengukur seberapa sering suatu kata muncul dalam satu dokumen. Semakin sering kata muncul, semakin besar bobotnya dalam dokumen tersebut. Pendekatan ini berangkat dari asumsi bahwa kata yang sering muncul cenderung lebih penting.

Secara matematis, TF dapat dihitung sebagai jumlah kemunculan kata dalam dokumen. Bentuk paling sederhana disebut frekuensi mentah. Jika kata "bagus" muncul tiga kali dalam satu dokumen, maka nilai TF mentahnya adalah tiga. Pendekatan ini mudah dipahami dan langsung digunakan dalam banyak sistem awal.

Namun frekuensi mentah memiliki kelemahan. Dokumen yang lebih panjang cenderung memiliki nilai frekuensi lebih tinggi. Kata yang sama dapat memiliki bobot berbeda hanya karena panjang dokumen berbeda. Untuk mengatasi hal ini, digunakan normalisasi frekuensi.

Normalisasi TF dilakukan dengan membagi jumlah kemunculan kata dengan total kata dalam dokumen. Rumus umum TF ter-normalisasi dapat dituliskan sebagai berikut.

$$TF(t,d) = \frac{f(t,d)}{\Sigma_{w \epsilon d} f(w,d)}$$

Nilai ini merepresentasikan proporsi kemunculan kata terhadap seluruh kata dalam dokumen. Dengan normalisasi, perbandingan antar dokumen menjadi lebih adil. Dokumen panjang dan pendek dapat dibandingkan dalam skala yang sama.

Perbedaan antara frekuensi mentah dan frekuensi relatif dapat dilihat pada contoh berikut.

Dokumen	Total Kata	Frekuensi "bagus"	TF Mentah	TF Relatif
D1	10	2	2	0,20
D2	100	2	2	0,02

Pada contoh tersebut, kedua dokumen memiliki frekuensi mentah yang sama. Namun bobot relatif berbeda karena panjang dokumen berbeda. Nilai relatif pada D1 lebih besar karena kata tersebut lebih dominan dalam dokumen pendek.

Panjang dokumen memengaruhi distribusi bobot secara signifikan. Tanpa normalisasi, dokumen panjang cenderung mendominasi proses pelatihan. Model dapat menganggap dokumen panjang lebih informatif meskipun isinya tidak

lebih relevan. Oleh karena itu, normalisasi TF menjadi langkah penting dalam sistem klasifikasi.

Meskipun sederhana, TF belum mempertimbangkan kelangkaan kata dalam korpus. Kata umum seperti "produk" dapat memiliki nilai tinggi pada banyak dokumen. Kondisi ini dapat mengurangi daya diskriminatif fitur. Untuk mengatasi masalah tersebut, TF sering dikombinasikan dengan Inverse Document Frequency dalam metode TF-IDF yang akan dibahas pada bagian berikutnya.

D. TF-IDF

TF-IDF merupakan pengembangan dari Term Frequency yang mempertimbangkan distribusi kata dalam seluruh korpus. Metode ini tidak hanya melihat frekuensi kata dalam satu dokumen, tetapi juga memperhatikan seberapa umum kata tersebut muncul pada dokumen lain. Dengan pendekatan ini, kata yang terlalu umum tidak akan mendapat bobot tinggi.

Konsep utama dalam TF-IDF adalah Inverse Document Frequency atau IDF. IDF mengukur tingkat kelangkaan suatu kata dalam korpus. Jika sebuah kata muncul di hampir semua dokumen, maka nilai IDF akan rendah. Sebaliknya, jika kata hanya muncul pada sedikit dokumen, nilai IDF akan tinggi. Prinsip ini membantu menonjolkan kata yang lebih spesifik.

Tujuan pembobotan kelangkaan kata adalah meningkatkan daya diskriminatif fitur. Kata umum seperti "dan" atau "produk" sering muncul di banyak dokumen. Jika hanya menggunakan TF, kata tersebut dapat mendominasi bobot meskipun tidak informatif. IDF menurunkan bobot kata yang terlalu sering muncul sehingga fitur menjadi lebih selektif.

Rumus matematis IDF dapat dituliskan sebagai berikut.

$$IDF(t, D) = log(\frac{N}{df(t)})$$

Keterangan:

N adalah jumlah dokumen dalam korpus.

df(*t*) adalah jumlah dokumen yang mengandung kata *t*.

Nilai TF-IDF diperoleh dengan mengalikan TF dan IDF.

$$TF - IDF(t, d, D) = TF(t, d) \times IDF\ (t, D)$$

Nilai TF-IDF tinggi menunjukkan bahwa kata tersebut sering muncul dalam dokumen tertentu tetapi jarang muncul dalam dokumen lain. Nilai rendah menunjukkan kata tersebut umum atau kurang relevan secara spesifik. Dengan demikian, TF-IDF membantu model fokus pada kata yang lebih informatif.

Interpretasi nilai TF-IDF dapat dipahami melalui contoh sederhana. Misalkan terdapat tiga dokumen:

D1: “produk bagus dan cepat”

D2: “produk mahal dan buruk”

D3: “produk bagus dan murah”

Jumlah dokumen N=3.

Kata “produk” muncul di ketiga dokumen sehingga df=3.

Kata “cepat” hanya muncul di satu dokumen sehingga df=1.

Nilai IDF untuk masing-masing kata dapat dihitung sebagai berikut.

Untuk “produk”:

$$IDF = log(\frac{3}{3}) = log(1) = 0$$

Untuk “cepat”:

$$IDF = log(\frac{3}{1}) = log(3)$$

Karena IDF “produk” bernilai nol, maka TF-IDF untuk kata tersebut juga akan rendah. Sebaliknya, kata “cepat” memiliki nilai IDF lebih tinggi sehingga bobotnya meningkat. Model akan lebih memperhatikan kata yang bersifat spesifik dibandingkan kata umum.

TF-IDF banyak digunakan dalam klasifikasi sentimen berbasis machine learning klasik. Metode ini relatif efisien dan mudah diinterpretasikan. Namun TF-IDF tetap tidak mempertimbangkan urutan kata dan konteks semantik. Oleh karena itu, pendekatan ini sering menjadi dasar sebelum menggunakan metode representasi yang lebih kompleks seperti embedding.

E. N-GRAM

Metode N-gram dikembangkan untuk mengatasi keterbatasan representasi unigram yang mengabaikan urutan kata. N-gram merepresentasikan teks sebagai rangkaian kata yang berurutan sepanjang N unit. Dengan cara ini, sistem dapat menangkap pola lokal dalam kalimat. Informasi urutan menjadi bagian dari fitur.

Unigram adalah representasi satu kata tunggal. Bigram terdiri dari dua kata yang berurutan. Trigram terdiri dari tiga kata yang berurutan. Semakin besar nilai N, semakin panjang konteks lokal yang ditangkap. Namun peningkatan nilai N juga meningkatkan kompleksitas fitur.

Pentingnya mempertahankan urutan lokal terlihat jelas dalam analisis sentimen. Kalimat “tidak bagus” memiliki makna berbeda dari “bagus tidak.” Jika hanya menggunakan unigram, kedua kalimat memiliki kata yang sama. Dengan bigram, pola “tidak bagus” dapat dikenali sebagai unit khusus. Pendekatan ini membantu sistem menangkap ekspresi negasi secara lebih akurat.

Pengaruh N-gram terhadap dimensi fitur cukup signifikan. Jika korpus memiliki seribu kata unik, jumlah unigram maksimal adalah seribu. Namun jumlah bigram dapat mencapai ratusan ribu tergantung kombinasi kata. Dimensi fitur meningkat secara eksponensial ketika nilai N bertambah. Kondisi ini memperbesar risiko sparsity.

Contoh sederhana dapat dilihat pada kalimat berikut.

Kalimat: “produk ini tidak bagus”

Unigram:

[produk, ini, tidak, bagus]

Bigram:

[produk ini, ini tidak, tidak bagus]

Trigram:

[produk ini tidak, ini tidak bagus]

Dari contoh tersebut terlihat bahwa bigram “tidak bagus” membawa informasi sentimen negatif yang lebih jelas. Jika hanya menggunakan unigram, sistem harus mengandalkan hubungan implisit antara “tidak” dan “bagus.” Dengan N-gram, hubungan tersebut menjadi eksplisit.

Perbandingan dampak N-gram terhadap dimensi dapat dilihat pada tabel berikut.

Nilai N	Jenis Representasi	Potensi Dimensi	Risiko Sparsity
1	Unigram	Rendah hingga sedang	Sedang
2	Bigram	Tinggi	Tinggi
3	Trigram	Sangat tinggi	Sangat tinggi

Kelebihan utama N-gram adalah kemampuannya menangkap konteks lokal dan pola frasa. Metode ini meningkatkan akurasi pada kasus negasi atau ekspresi tetap. N-gram juga relatif mudah diintegrasikan dengan TF atau TF-IDF.

Namun risiko sparsity meningkat seiring bertambahnya nilai N. Banyak kombinasi kata jarang muncul dalam korpus. Matriks fitur menjadi sangat jarang dan berdimensi besar. Model memerlukan memori dan waktu pelatihan lebih besar. Oleh karena itu, pemilihan nilai N perlu mempertimbangkan ukuran dataset dan tujuan penelitian.

N-gram menjadi jembatan antara representasi frekuensi sederhana dan embedding berbasis konteks. Metode ini tetap relevan dalam banyak studi analisis sentimen, terutama ketika fokus pada ekspresi frasa pendek dan negasi.

F. WORD EMBEDDING

Word embedding memperkenalkan pendekatan baru dalam representasi teks. Metode ini tidak lagi menghitung frekuensi kata secara langsung. Sebaliknya, setiap kata dipetakan ke dalam vektor berdimensi rendah yang padat. Representasi ini dibangun berdasarkan konteks kemunculan kata dalam korpus besar.

Ilustrasi Kemunculan Kata pada Analisis Sentimen

Konsep dasar embedding berangkat dari hipotesis distribusional. Kata yang muncul dalam konteks serupa cenderung memiliki makna yang mirip. Jika kata “bagus” sering muncul bersama kata “pelayanan” dan “kualitas,” maka vektornya akan berada dekat dengan kata evaluatif lain. Pendekatan ini memungkinkan sistem menangkap relasi semantik secara otomatis.

Perbedaan utama antara dense dan sparse vector terletak pada jumlah nilai nol. Representasi seperti Bag of Words menghasilkan vektor berdimensi sangat besar dengan banyak nilai nol. Kondisi ini disebut sparse. Word embedding menghasilkan vektor berdimensi kecil dengan hampir semua nilai terisi angka. Kondisi ini disebut dense. Dense vector lebih efisien dalam penyimpanan dan lebih kaya informasi semantik.

Word2Vec merupakan salah satu metode embedding yang populer. Model ini dikembangkan untuk mempelajari representasi kata melalui jaringan saraf sederhana. Terdapat dua arsitektur utama dalam Word2Vec, yaitu Continuous Bag of Words dan Skip-gram. CBOW memprediksi kata target berdasarkan konteks di sekitarnya. Skip-gram melakukan kebalikan, yaitu

memprediksi konteks berdasarkan kata target. Kedua pendekatan ini menghasilkan vektor kata yang mencerminkan hubungan semantik.

GloVe atau Global Vectors menggunakan pendekatan berbeda. Model ini memanfaatkan matriks ko-occurence global dari seluruh korpus. Setiap elemen matriks mencerminkan seberapa sering dua kata muncul bersama. GloVe mengoptimalkan representasi vektor agar dapat merekonstruksi statistik kemunculan tersebut. Pendekatan ini menggabungkan keunggulan informasi lokal dan global.

FastText mengembangkan embedding dengan mempertimbangkan subword atau potongan karakter. Kata dipecah menjadi beberapa unit karakter kecil. Dengan cara ini, model dapat menangkap struktur morfologis bahasa. Pendekatan ini efektif untuk bahasa dengan banyak imbuhan seperti bahasa Indonesia. FastText juga membantu menangani kata yang jarang muncul atau belum pernah muncul sebelumnya.

Perbandingan singkat metode embedding dapat dilihat pada tabel berikut.

Metode	Prinsip Dasar	Kelebihan	Keterbatasan
Word2Vec	Prediksi konteks	Efisien dan populer	Tidak tangkap konteks kalimat penuh
GloVe	Ko-occurence global	Stabil secara statistik	Perlu korpus besar
FastText	Representasi subword	Baik untuk bahasa berimbuhan	Dimensi bisa lebih besar

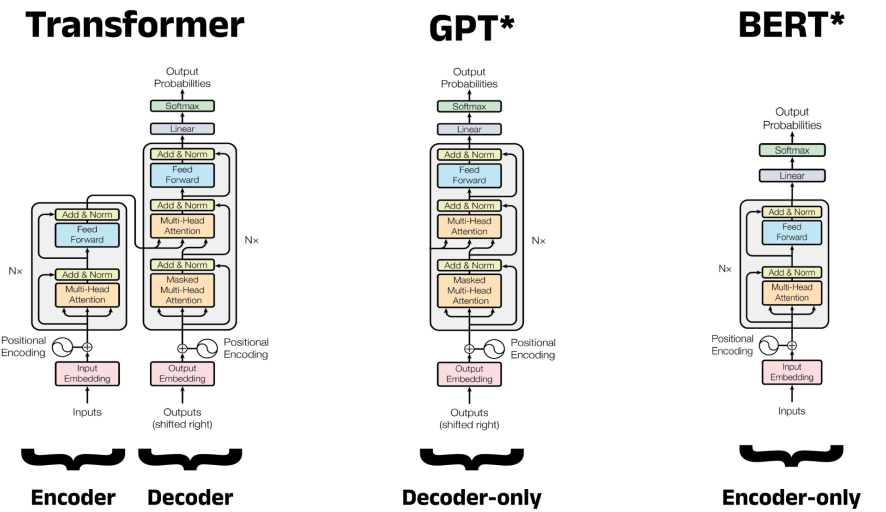
Kelebihan utama embedding terletak pada kemampuannya menangkap makna semantik. Kata yang memiliki arti serupa akan memiliki jarak vektor yang dekat. Relasi seperti analogi juga dapat dipelajari secara otomatis. Sebagai contoh, hubungan antara “raja” dan “ratu” dapat tercermin dalam struktur vektor.

Dalam analisis sentimen, embedding membantu model memahami konteks evaluatif secara lebih dalam. Sinonim dan variasi kata dapat dipetakan dalam ruang yang sama. Hal ini mengurangi masalah sparsity dan meningkatkan generalisasi model. Word embedding menjadi langkah penting menuju representasi kontekstual yang lebih kompleks pada bagian berikutnya.

G. CONTEXTUAL EMBEDDING

Word embedding seperti Word2Vec dan GloVe bersifat statis. Setiap kata memiliki satu vektor tetap tanpa memperhatikan konteks kalimat. Kata "ringan" akan memiliki representasi yang sama dalam kalimat "tas ini ringan" dan "hukuman itu ringan." Padahal maknanya berbeda. Keterbatasan ini menjadi kendala dalam analisis sentimen yang sensitif terhadap konteks.

Embedding kontekstual hadir untuk mengatasi masalah tersebut. Representasi ini dibangun menggunakan arsitektur transformer. Model transformer menggunakan mekanisme perhatian untuk membaca seluruh kalimat secara simultan. Setiap kata memperoleh vektor yang bergantung pada kata lain di sekitarnya. Dengan cara ini, makna kata menjadi dinamis.



Arsitektur Transformer

Salah satu model yang paling dikenal adalah BERT. Model ini menggunakan pendekatan bidirectional. Artinya, model membaca konteks dari kiri dan kanan sekaligus. Proses pelatihan awal dilakukan pada korpus besar melalui tugas prediksi kata tersembunyi. Setelah pelatihan awal selesai, model dapat disesuaikan untuk berbagai tugas termasuk analisis sentimen.

Arsitektur dasar BERT terdiri dari beberapa lapisan encoder transformer. Setiap lapisan memiliki mekanisme self-attention dan feed-forward network. Self-attention menghitung hubungan antar kata dalam satu kalimat.

Proses ini memungkinkan model memahami dependensi jarak jauh. Hasil akhirnya adalah representasi kontekstual untuk setiap token.

Fine-tuning menjadi langkah penting dalam penerapan BERT pada analisis sentimen. Model pralatih digunakan sebagai dasar, kemudian dilatih kembali dengan dataset berlabel sentimen. Proses ini menyesuaikan bobot model terhadap tugas spesifik. Fine-tuning biasanya membutuhkan data lebih sedikit dibanding pelatihan dari awal. Pendekatan ini meningkatkan efisiensi dan akurasi.

Perbandingan embedding statis dan kontekstual dapat dilihat pada tabel berikut.

Aspek	Embedding Statis	Embedding Kontekstual
Representasi kata	Tetap	Bergantung konteks
Tangkap makna ambigu	Terbatas	Lebih baik
Kompleksitas komputasi	Sedang	Tinggi
Kebutuhan data	Relatif besar	Lebih efisien melalui fine-tuning
Performa analisis sentimen	Baik	Umumnya lebih tinggi

Embedding kontekstual memberikan pemahaman makna yang lebih dalam. Model mampu membedakan penggunaan kata dalam situasi berbeda. Hal ini penting ketika kalimat mengandung ironi, negasi, atau makna ganda. Namun kebutuhan komputasi dan sumber daya meningkat secara signifikan.

Perkembangan embedding kontekstual mengubah pendekatan analisis sentimen secara mendasar. Representasi tidak lagi sekadar frekuensi atau asosiasi global. Model kini mempertimbangkan konteks penuh dalam satu kalimat. Pendekatan ini membuka peluang akurasi lebih tinggi, tetapi juga menuntut perencanaan sumber daya yang matang.

H. SPARSE VECTOR VS DENSE VECTOR

Representasi fitur dalam analisis teks umumnya terbagi menjadi dua kategori besar, yaitu sparse dan dense. Perbedaan keduanya terletak pada struktur vektor dan distribusi nilainya. Pemahaman terhadap perbedaan ini penting karena berdampak langsung pada efisiensi komputasi dan performa model.

Sparse representation adalah representasi dengan banyak nilai nol dalam vektornya. Metode seperti Bag of Words dan TF-IDF menghasilkan vektor berdimensi sangat tinggi. Setiap kata unik menjadi satu dimensi. Karena satu dokumen hanya mengandung sebagian kecil dari seluruh kosakata, sebagian besar nilai akan bernilai nol. Kondisi ini disebut sparsity.

Dense representation adalah representasi dengan dimensi lebih rendah dan sedikit nilai nol. Word embedding dan contextual embedding termasuk dalam kategori ini. Setiap kata direpresentasikan dalam ruang berdimensi tetap, misalnya 100 atau 300 dimensi. Hampir seluruh elemen vektor memiliki nilai non-nol. Representasi ini lebih padat dan informatif.

Dampak terhadap penyimpanan cukup signifikan. Sparse vector membutuhkan memori besar jika disimpan secara penuh. Namun teknik kompresi khusus dapat mengurangi beban penyimpanan. Dense vector memiliki dimensi lebih kecil, tetapi setiap elemen harus disimpan karena bernilai penting. Dalam skala besar, embedding juga memerlukan kapasitas penyimpanan yang besar.

Dari sisi komputasi, sparse vector relatif ringan untuk model linear. Algoritma seperti Naive Bayes dan Logistic Regression dapat memproses matriks jarang secara efisien. Dense vector lebih cocok untuk deep learning yang memanfaatkan operasi matriks intensif. Namun pelatihan model berbasis embedding membutuhkan perangkat keras yang lebih kuat.

Kinerja model juga berbeda tergantung pendekatan yang digunakan. Pada machine learning klasik dengan dataset terbatas, TF-IDF sering memberikan hasil yang stabil. Pada deep learning dengan dataset besar, embedding kontekstual umumnya menghasilkan akurasi lebih tinggi. Pilihan representasi perlu disesuaikan dengan tujuan penelitian dan sumber daya yang tersedia.

Perbandingan ringkas antara sparse dan dense vector dapat dilihat pada tabel berikut.

Aspek	Sparse Vector (BoW, TF-IDF)	Dense Vector (Embedding)
Dimensi	Sangat tinggi	Rendah
Nilai nol	Banyak	Sedikit
Tangkap makna semantik	Terbatas	Lebih baik
Kebutuhan komputasi	Lebih ringan	Lebih berat

Perbandingan ini menunjukkan bahwa tidak ada representasi yang selalu unggul. Sparse vector lebih sederhana dan mudah diinterpretasikan. Dense vector lebih kaya secara semantik dan adaptif terhadap konteks. Pemilihan metode harus mempertimbangkan keseimbangan antara akurasi, efisiensi, dan kebutuhan penelitian.

I. PEMILIHAN REPRESENTASI

Pemilihan representasi fitur tidak dapat dilakukan secara sembarangan. Setiap metode memiliki karakteristik dan konsekuensi teknis. Peneliti perlu menyesuaikan pilihan dengan tujuan penelitian dan kondisi data. Keputusan ini memengaruhi akurasi, waktu komputasi, dan interpretasi hasil.

Ukuran dataset menjadi pertimbangan awal. Pada dataset kecil, representasi sederhana seperti TF-IDF sering lebih stabil. Model linear dapat belajar dengan baik tanpa membutuhkan parameter besar. Embedding kontekstual cenderung membutuhkan data lebih banyak agar tidak overfitting. Jika data terbatas, model kompleks dapat menghasilkan varians tinggi.

Pada dataset besar, pendekatan berbasis embedding lebih efektif. Model deep learning mampu menangkap pola semantik yang lebih kompleks. Dengan data cukup, representasi kontekstual dapat meningkatkan generalisasi. Namun kebutuhan komputasi juga meningkat secara signifikan. Peneliti perlu memastikan ketersediaan sumber daya sebelum memilih metode tersebut.

Jenis model juga menentukan pilihan representasi. Model klasik seperti Naive Bayes dan Support Vector Machine bekerja baik dengan fitur sparse. Representasi TF-IDF sering menjadi pilihan utama pada skenario ini. Sebaliknya, model deep learning memerlukan dense vector sebagai input. Embedding statis atau kontekstual lebih sesuai untuk arsitektur neural.

Aspek interpretabilitas juga perlu dipertimbangkan. Representasi berbasis frekuensi lebih mudah dijelaskan. Peneliti dapat menelusuri kata mana yang berkontribusi terhadap prediksi. Embedding dan transformer menghasilkan representasi yang sulit diinterpretasikan secara langsung. Jika penelitian menekankan transparansi, metode sederhana mungkin lebih sesuai.

Performa sering menjadi alasan utama memilih model kompleks. Embedding kontekstual biasanya memberikan akurasi lebih tinggi pada tugas analisis sentimen yang sulit. Namun peningkatan performa tidak selalu signifikan pada semua dataset. Dalam beberapa kasus, perbedaan akurasi hanya kecil tetapi biaya komputasi jauh lebih besar.

Keseimbangan antara kompleksitas dan efisiensi menjadi prinsip penting. Representasi yang terlalu sederhana dapat kehilangan informasi penting. Representasi yang terlalu kompleks dapat membebani sistem tanpa peningkatan berarti. Peneliti perlu melakukan evaluasi empiris untuk menentukan pilihan yang paling sesuai.

Ringkasan pertimbangan pemilihan representasi dapat dilihat pada tabel berikut.

Kondisi	Representasi yang Disarankan
Dataset kecil	TF atau TF-IDF
Dataset besar	Word Embedding atau Transformer
Model klasik	Sparse vector
Deep learning	Dense vector
Fokus interpretabilitas	BoW atau TF-IDF
Fokus akurasi tinggi	Embedding kontekstual

Pemilihan representasi bukan keputusan tunggal yang bersifat absolut. Setiap penelitian memiliki konteks dan keterbatasan berbeda. Pendekatan yang tepat adalah menyesuaikan metode dengan tujuan, data, dan sumber daya yang tersedia. Dengan pertimbangan tersebut, sistem analisis sentimen dapat dirancang secara lebih efektif dan rasional.

BAB 7. MODEL MACHINE LEARNING UNTUK ANALISIS SENTIMEN

Muhammad Anwar Fauzi, S.Kom., M.Kom.

A. KONSEP DASAR MACHINE LEARNING

Machine learning dalam analisis sentimen umumnya menggunakan pendekatan supervised learning. Pada pendekatan ini, sistem belajar dari data yang telah diberi label. Label biasanya berupa kategori seperti positif, negatif, atau netral. Model mempelajari hubungan antara fitur teks dan label tersebut.

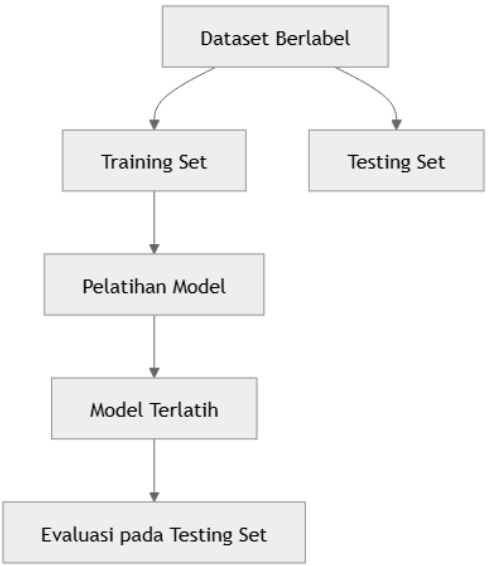
Klasifikasi teks menjadi tugas utama dalam analisis sentimen berbasis machine learning. Setiap dokumen direpresentasikan sebagai vektor fitur. Vektor ini kemudian menjadi input bagi algoritma klasifikasi. Model mencoba menemukan pola yang dapat memisahkan kelas secara konsisten.

Dataset berlabel menjadi komponen penting dalam proses pelatihan. Data ini biasanya dibagi menjadi dua bagian utama, yaitu data latih dan data uji. Data latih digunakan untuk membangun model. Data uji digunakan untuk mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya. Pembagian ini membantu mengukur kemampuan generalisasi.

Representasi fitur berperan sebagai jembatan antara teks dan algoritma. Metode seperti TF-IDF menghasilkan matriks numerik yang dapat diproses model. Kualitas fitur sangat mempengaruhi hasil klasifikasi. Fitur yang tidak relevan dapat menurunkan akurasi dan meningkatkan kompleksitas.

Dalam praktik, dataset sering dibagi menjadi tiga bagian, yaitu training set, validation set, dan testing set. Training set digunakan untuk melatih model. Validation set digunakan untuk menyetel parameter dan memilih model terbaik. Testing set digunakan untuk evaluasi akhir setelah semua penyesuaian selesai.

Alur umum pelatihan model dapat digambarkan sebagai berikut.



Alur Pelatihan Model

Metrik evaluasi dasar digunakan untuk mengukur performa model. Akurasi mengukur proporsi prediksi yang benar. Precision mengukur ketepatan prediksi pada kelas tertentu. Recall mengukur kemampuan model mendeteksi seluruh contoh pada kelas tersebut. F1-score menggabungkan precision dan recall dalam satu nilai harmonis.

Pemilihan metrik bergantung pada tujuan penelitian. Pada dataset tidak seimbang, akurasi saja tidak cukup. Model dapat mencapai akurasi tinggi dengan hanya memprediksi kelas mayoritas. Oleh karena itu, precision, recall, dan F1-score sering digunakan secara bersamaan.

Pemahaman konsep dasar ini menjadi fondasi sebelum mempelajari algoritma spesifik. Setiap model memiliki asumsi dan karakteristik berbeda. Namun semua model mengikuti kerangka umum supervised learning dalam klasifikasi teks. Dengan fondasi ini, pembahasan dapat dilanjutkan pada algoritma pertama yang banyak digunakan dalam analisis sentimen, yaitu Naïve Bayes.

B. NAÏVE BAYES

Naïve Bayes merupakan salah satu algoritma paling awal yang digunakan dalam klasifikasi teks. Model ini berbasis probabilistik dan berakar pada Teorema Bayes. Pendekatan ini menghitung probabilitas suatu dokumen termasuk dalam kelas tertentu berdasarkan fitur yang dimilikinya. Dalam analisis sentimen, kelas biasanya berupa positif atau negatif.

Teorema Bayes dapat dituliskan sebagai berikut.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Keterangan:
c adalah kelas sentimen.
d adalah dokumen.
$P(c|d)$ adalah probabilitas kelas setelah melihat dokumen.
$P(d|c)$ adalah probabilitas dokumen muncul pada kelas tersebut.
$P(c)$ adalah probabilitas awal kelas.

Model memilih kelas dengan probabilitas posterior tertinggi. Dalam praktik, penyebut $P(d)$ sering diabaikan karena nilainya sama untuk semua kelas. Proses klasifikasi menjadi perbandingan nilai numerik antar kelas.

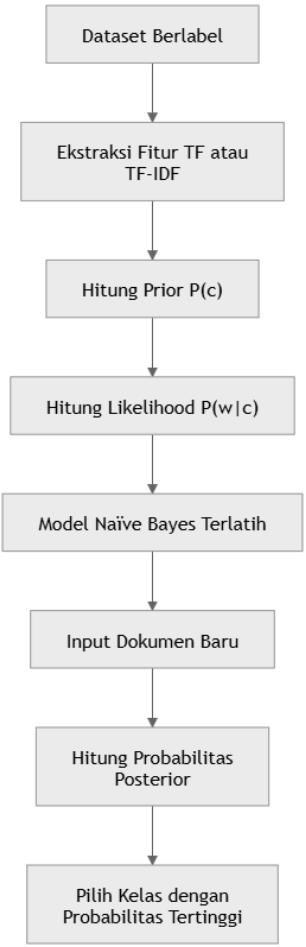
Istilah "naïve" merujuk pada asumsi independensi fitur. Model mengasumsikan bahwa setiap kata dalam dokumen bersifat independen terhadap kata lain. Artinya, kemunculan satu kata tidak memengaruhi probabilitas kemunculan kata lain. Asumsi ini jarang sepenuhnya benar dalam bahasa alami. Namun dalam banyak kasus, model tetap memberikan hasil yang cukup baik.

Untuk teks, varian yang paling umum adalah Multinomial Naïve Bayes. Model ini menggunakan frekuensi kata sebagai dasar perhitungan probabilitas. Setiap kata dihitung berdasarkan kemunculannya dalam dokumen dan dalam kelas tertentu. Probabilitas dihitung menggunakan distribusi multinomial.

Dalam implementasi, digunakan teknik smoothing seperti Laplace smoothing untuk menghindari probabilitas nol. Tanpa smoothing, kata yang tidak pernah muncul dalam kelas tertentu akan menghasilkan probabilitas nol dan mengganggu perhitungan.

Naïve Bayes memiliki kelebihan pada data berdimensi tinggi. Representasi teks seperti TF-IDF sering menghasilkan ribuan fitur. Model ini tetap efisien karena perhitungannya sederhana. Proses pelatihan relatif cepat dibandingkan model kompleks. Algoritma ini juga bekerja baik pada dataset kecil hingga menengah.

Namun keterbatasan utama terletak pada asumsi independensi fitur. Dalam bahasa alami, kata sering saling berkaitan. Frasa seperti "tidak bagus" memiliki makna khusus yang tidak dapat ditangkap jika kata diperlakukan independen. Model juga kurang efektif dalam menangani hubungan konteks yang kompleks.



Alur Naive Bayes

Meskipun memiliki keterbatasan, Naïve Bayes tetap menjadi baseline yang kuat dalam analisis sentimen. Model ini sering digunakan sebagai pembanding sebelum menerapkan algoritma yang lebih kompleks. Kesederhanaan dan efisiensinya membuatnya relevan dalam berbagai studi klasifikasi teks.

C. LOGISTIC REGRESSION

Logistic Regression merupakan model linear yang banyak digunakan dalam klasifikasi biner. Dalam analisis sentimen, model ini sering digunakan untuk membedakan kelas positif dan negatif. Meskipun namanya mengandung

istilah regression, metode ini digunakan untuk tugas klasifikasi. Model memprediksi probabilitas suatu dokumen termasuk dalam kelas tertentu.

Logistic Regression membangun kombinasi linear dari seluruh fitur. Setiap fitur memiliki bobot atau koefisien. Nilai kombinasi linear tersebut kemudian dipetakan ke rentang 0 hingga 1 menggunakan fungsi sigmoid. Fungsi sigmoid dapat dituliskan sebagai berikut.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Nilai z merupakan hasil perkalian antara vektor fitur dan vektor bobot. Output sigmoid diinterpretasikan sebagai probabilitas kelas positif. Jika probabilitas lebih besar dari ambang tertentu, dokumen diklasifikasikan sebagai positif. Jika tidak, dokumen diklasifikasikan sebagai negatif.

Keunggulan utama model ini terletak pada interpretasi koefisien fitur. Setiap koefisien menunjukkan arah dan kekuatan pengaruh suatu kata terhadap kelas. Koefisien positif menunjukkan bahwa kata tersebut meningkatkan kemungkinan kelas positif. Koefisien negatif menunjukkan kecenderungan ke arah negatif. Interpretasi ini membantu peneliti memahami pola sentimen dalam data.

Pada data teks berdimensi tinggi, risiko overfitting cukup besar. Logistic Regression mengatasi masalah ini melalui regularization. Regularization menambahkan penalti pada besar koefisien agar model tidak terlalu menyesuaikan diri dengan data latih. Dua jenis regularization yang umum digunakan adalah L1 dan L2.

Regularization L1 menambahkan penalti berdasarkan nilai absolut koefisien. Pendekatan ini mendorong sebagian koefisien menjadi nol. Hasilnya adalah model yang lebih sederhana dan seleksi fitur otomatis. Regularization L2 menambahkan penalti berdasarkan kuadrat koefisien. Pendekatan ini mengecilkan bobot tanpa menghilangkan fitur sepenuhnya. Kedua teknik membantu meningkatkan generalisasi model.

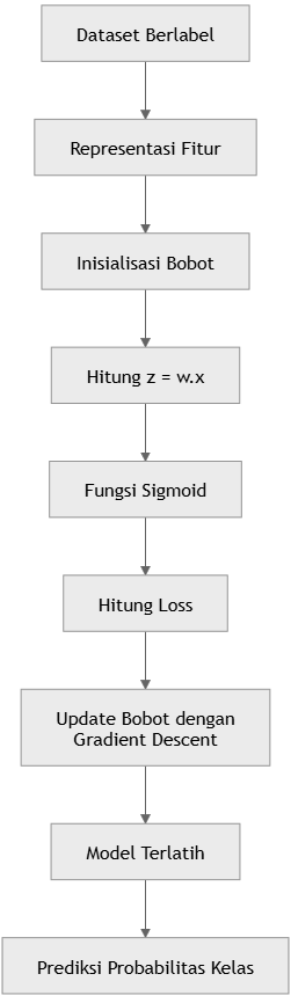


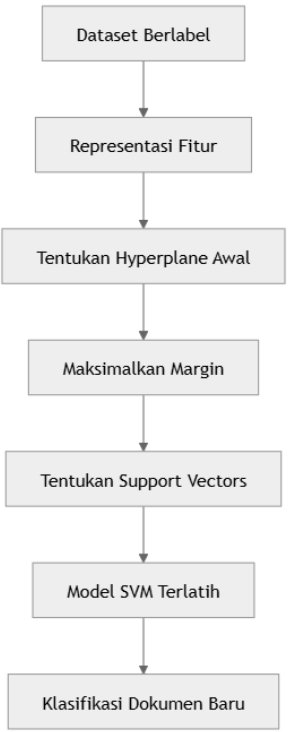
Diagram Alir Logistic Regression

Dari sisi interpretabilitas, Logistic Regression lebih transparan dibandingkan banyak model lain. Peneliti dapat mengidentifikasi kata mana yang paling berkontribusi pada prediksi. Hal ini penting dalam penelitian akademik yang menuntut penjelasan rasional. Model ini juga relatif efisien dan stabil pada data teks dengan representasi TF-IDF.

Namun Logistic Regression tetap merupakan model linear. Model ini kurang mampu menangkap hubungan non-linear antar fitur. Jika pola sentimen sangat kompleks, performa dapat terbatas. Meskipun demikian, keseimbangan antara akurasi dan interpretabilitas menjadikan Logistic Regression salah satu algoritma yang paling sering digunakan dalam analisis sentimen

D. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine merupakan algoritma klasifikasi yang bekerja dengan mencari batas pemisah optimal antar kelas. Batas ini disebut hyperplane. Dalam analisis sentimen, hyperplane memisahkan dokumen positif dan negatif dalam ruang fitur. Tujuan utama model adalah memaksimalkan jarak antara hyperplane dan titik data terdekat dari masing-masing kelas.



Alur Proses SVM

Konsep margin menjadi inti dalam SVM. Margin adalah jarak antara hyperplane dan titik data terdekat yang disebut support vectors. Model berusaha memilih hyperplane dengan margin terbesar. Margin yang besar membantu meningkatkan kemampuan generalisasi. Dengan cara ini, model tidak hanya memisahkan data latih, tetapi juga lebih stabil pada data baru.

Untuk data teks, SVM linear sering digunakan. Representasi seperti TF-IDF menghasilkan ruang fitur berdimensi tinggi. SVM linear bekerja efektif dalam kondisi ini karena mampu menangani banyak fitur. Algoritma ini juga relatif tahan terhadap overfitting pada data sparse. Oleh karena itu, SVM sering memberikan performa kuat pada klasifikasi sentimen.

Pada kasus di mana data tidak dapat dipisahkan secara linear, digunakan kernel trick. Kernel memungkinkan pemetaan data ke ruang dimensi lebih tinggi tanpa menghitung transformasi secara eksplisit. Kernel yang umum digunakan meliputi linear, polynomial, dan radial basis function. Namun dalam klasifikasi teks, kernel linear sering sudah memadai.

Keunggulan utama SVM terletak pada akurasi dan stabilitasnya. Model ini sering menjadi pilihan utama dalam kompetisi klasifikasi teks sebelum era deep learning. SVM juga efektif ketika jumlah fitur jauh lebih besar dibanding jumlah sampel. Hal ini umum terjadi pada data teks.

Namun SVM memiliki beberapa keterbatasan. Pemilihan parameter seperti nilai C dan jenis kernel memengaruhi performa secara signifikan. Proses pelatihan dapat menjadi lambat pada dataset sangat besar. Selain itu, interpretasi model tidak semudah Logistic Regression karena tidak langsung memberikan probabilitas kelas tanpa modifikasi tambahan.

Secara umum, SVM menawarkan keseimbangan antara kekuatan klasifikasi dan ketahanan terhadap dimensi tinggi. Dalam banyak penelitian analisis sentimen berbasis machine learning klasik, SVM sering digunakan sebagai model pembanding utama. Model ini tetap relevan terutama ketika data tersedia dalam jumlah sedang dan representasi fitur bersifat sparse.

E. K-NEAREST NEIGHBOR (KNN)

K-Nearest Neighbor merupakan algoritma klasifikasi berbasis kemiripan. Model ini tidak membangun parameter eksplisit selama pelatihan. Sebaliknya, model menyimpan seluruh data latih. Ketika dokumen baru masuk, sistem menghitung jaraknya terhadap seluruh dokumen latih. Kelas ditentukan berdasarkan mayoritas tetangga terdekat.

Konsep utama KNN adalah jarak dalam ruang fitur. Jarak sering dihitung menggunakan Euclidean distance atau cosine similarity. Pada data teks dengan representasi TF-IDF, cosine similarity lebih umum digunakan. Metode ini mengukur sudut antar vektor sehingga lebih stabil pada data berdimensi tinggi.

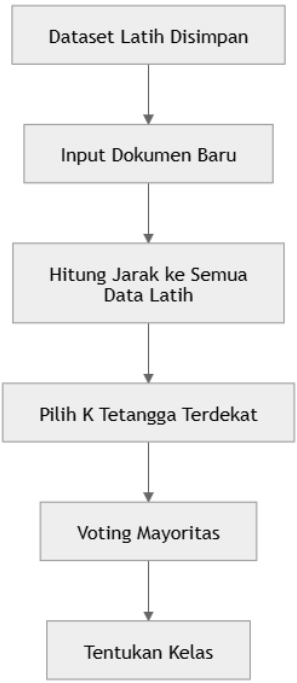


Diagram Alur KNN

Nilai K menentukan jumlah tetangga yang dipertimbangkan. Jika K terlalu kecil, model menjadi sensitif terhadap noise. Jika K terlalu besar, batas antar kelas menjadi kabur. Pemilihan nilai K biasanya dilakukan melalui validasi silang. Nilai optimal bergantung pada distribusi data.

KNN memiliki kelebihan dalam kesederhanaan konsep. Algoritma ini tidak memerlukan asumsi distribusi data. Model juga dapat menangani klasifikasi multi-kelas tanpa modifikasi besar. Namun kompleksitas komputasi cukup tinggi pada tahap prediksi. Setiap prediksi memerlukan perhitungan jarak terhadap seluruh data latih.

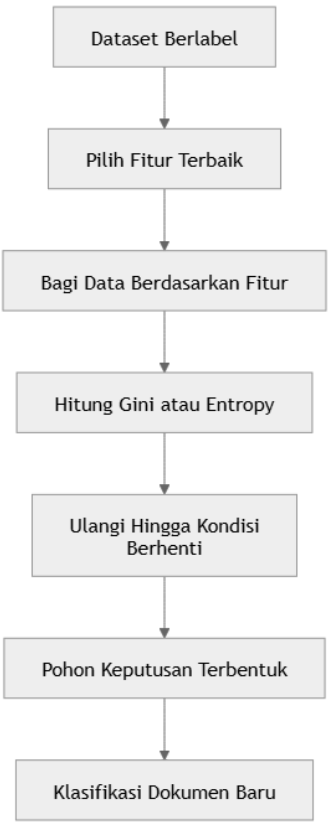
Pada data teks berdimensi tinggi, KNN menghadapi tantangan serius. Fenomena curse of dimensionality dapat mengurangi makna jarak. Ketika dimensi sangat besar, jarak antar dokumen cenderung menjadi seragam. Kondisi ini menurunkan kemampuan model membedakan kelas secara efektif.

Secara umum, KNN jarang menjadi pilihan utama pada analisis sentimen berskala besar. Namun algoritma ini tetap berguna sebagai pembanding sederhana. Dalam dataset kecil dengan fitur terkontrol, KNN dapat memberikan

hasil yang cukup baik. Pemahaman terhadap prinsip jarak dan kemiripan juga membantu menjelaskan dasar klasifikasi berbasis vektor.

F. DECISION TREE

Decision Tree membangun model dalam bentuk struktur pohon. Setiap node mewakili keputusan berdasarkan fitur tertentu. Cabang mewakili hasil keputusan tersebut. Daun pohon menunjukkan kelas akhir. Model ini bekerja dengan membagi data secara bertahap hingga mencapai pemisahan yang optimal.



Cara Kerja Decision Tree

Pemilihan fitur pada setiap node biasanya menggunakan kriteria seperti Gini impurity atau entropy. Kedua ukuran ini menghitung tingkat ketidakmurnian dalam subset data. Model memilih fitur yang paling efektif memisahkan kelas. Proses ini berlangsung secara rekursif hingga kondisi berhenti terpenuhi.

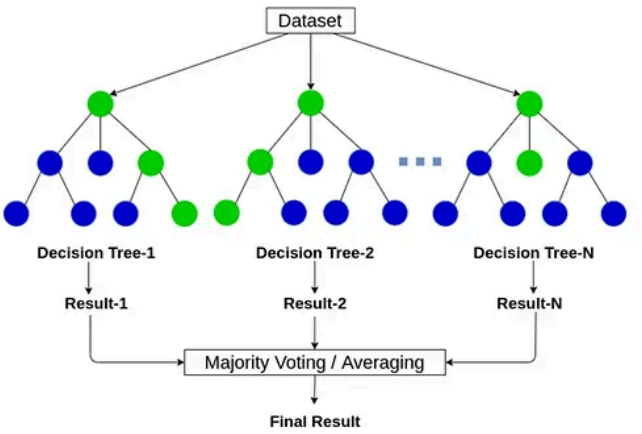
Keunggulan utama Decision Tree adalah interpretabilitasnya. Struktur pohon dapat divisualisasikan dan dipahami dengan mudah. Peneliti dapat melihat jalur keputusan dari akar hingga daun. Transparansi ini penting dalam penelitian yang menuntut penjelasan model.

Namun Decision Tree rentan terhadap overfitting. Pohon yang terlalu dalam dapat menyesuaikan diri secara berlebihan pada data latih. Model menjadi sangat spesifik dan kehilangan kemampuan generalisasi. Teknik seperti pruning digunakan untuk mengurangi kompleksitas pohon.

Pada data teks dengan dimensi tinggi, Decision Tree kurang populer dibandingkan model linear. Jumlah fitur yang sangat banyak dapat membuat pohon menjadi besar dan tidak stabil. Meskipun demikian, model ini tetap berguna untuk analisis eksploratif dan interpretasi awal.

G. RANDOM FOREST

Random Forest merupakan metode ensemble yang menggabungkan banyak Decision Tree. Setiap pohon dilatih pada subset data yang berbeda melalui teknik bagging. Selain itu, setiap node hanya mempertimbangkan subset fitur secara acak. Pendekatan ini meningkatkan variasi antar pohon.



Prediksi akhir diperoleh melalui voting mayoritas dari seluruh pohon. Pendekatan ini meningkatkan stabilitas dan akurasi dibandingkan satu pohon tunggal. Random Forest juga lebih tahan terhadap overfitting. Kombinasi banyak model mengurangi risiko kesalahan spesifik pada satu pohon.

Diagram Alir Random Forest

Model ini juga menyediakan informasi feature importance. Fitur yang sering digunakan dalam pemisahan akan memiliki nilai penting lebih tinggi. Informasi ini membantu peneliti memahami kontribusi fitur dalam klasifikasi.

Namun Random Forest tetap memiliki keterbatasan. Interpretasi tidak sesederhana Decision Tree tunggal. Proses pelatihan juga lebih berat secara komputasi. Pada data teks berdimensi sangat tinggi, performa dapat bergantung pada seleksi fitur sebelumnya.

Secara umum, Random Forest menawarkan keseimbangan antara akurasi dan stabilitas. Model ini sering digunakan ketika peneliti menginginkan performa lebih tinggi dibanding Decision Tree tanpa beralih ke deep learning.

BAB 8. MODEL DEEP LEARNING UNTUK ANALISIS SENTIMEN

Satria Pradana Rizki Yulianto, M.Kom.

A. KONSEP DEEP LEARNING

Sebelum dominasi Deep Learning, analisis sentimen sangat bergantung pada algoritma Machine Learning (ML) tradisional seperti Support Vector Machines (SVM), Naive Bayes, dan Decision Trees (Vidya & Janani, 2022). Meskipun algoritma ini memberikan dasar yang kuat, mereka memiliki keterbatasan fundamental yang dikenal sebagai ketergantungan pada feature engineering manual.

Dalam paradigma ML tradisional, efektivitas model sangat bergantung pada kualitas fitur yang diekstraksi oleh manusia. Pengguna harus secara manual menentukan atribut apa yang "penting" bagi model, seperti frekuensi kata (TF-IDF), keberadaan tanda baca tertentu, atau penggunaan leksikon sentimen (kamus kata positif/negatif). Proses ini memakan waktu, rentan terhadap bias manusia, dan sering kali gagal menangkap nuansa kontekstual yang kompleks (Lagrari & Elkettani, 2021).

Deep Learning mengatasi hambatan feature engineering melalui konsep representation learning. Model DL, yang terdiri dari lapisan-lapisan jaringan saraf tiruan (neural networks), memiliki kemampuan untuk mengekstrak fitur secara otomatis dari data mentah. Lapisan awal mungkin mendeteksi pola karakter sederhana, lapisan tengah mendeteksi kata atau frasa, dan lapisan akhir memahami struktur kalimat dan semantik yang kompleks.

Penelitian menunjukkan bahwa meskipun model ML tradisional lebih mudah diinterpretasikan (kita tahu mengapa keputusan dibuat berdasarkan fitur yang kita buat), model DL secara konsisten mengungguli dalam hal akurasi pada dataset besar karena kemampuannya menangkap hubungan non-linear dan dependensi jarak jauh dalam teks. Namun, kekuatan ini datang dengan biaya komputasi yang lebih tinggi dan kebutuhan akan perangkat keras khusus seperti GPU (Poomka et al., 2021).

Kalian bisa bayangkan ML tradisional itu layaknya navigasi menggunakan peta manual. Sebelum perjalanan dimulai, pengguna (manusia) harus menandai rute, menghitung jarak, dan memprediksi belokan penting (feature engineering). Jika ada jalan baru atau kemacetan mendadak (pola data baru yang tidak terduga), peta tersebut tidak akan membantu, karena pengguna tidak memasukkan informasi itu sebelumnya. Pengguna harus memberi tahu sistem secara eksplisit "Belok kiri jika melihat pohon besar."

Sebaliknya, Deep Learning berfungsi seperti sistem GPS modern yang terhubung satelit. Pengguna tinggal memberikan tujuan akhir (label sentimen: positif/negatif", dan sistem secara otomatis memindai seluruh medan, mempelajari pola lalu lintas (fitur data), dan menentukan rute terbaik tanpa intervensi manusia di setiap belokan.

B. FONDASI DEEP LEARNING

Untuk memahami bagaimana Deep Learning memproses bahasa, kita harus terlebih dahulu memahami bagaimana kata-kata diubah menjadi format yang dapat diproses oleh mesin. Komputer tidak memahami makna intrinsik dari kata "bahagia" atau "sedih"; mereka hanya memproses angka.

1. Word Embeddings

Teknologi kunci yang memungkinkan lompatan besar dalam NLP adalah Word Embeddings (seperti Word2Vec, GloVe, dan FastText). Berbeda dengan metode lama One-Hot Encoding yang merepresentasikan kata sebagai angka biner yang terisolasi dan jarang (sparse), Word Embeddings memetakan kata ke dalam vektor bernilai riil dalam ruang multidimensi (turing.com, 2022). Dalam ruang vektor ini, posisi geometris merepresentasikan makna semantik. Kata-kata yang memiliki makna serupa atau sering muncul dalam konteks yang sama akan ditempatkan berdekatan satu sama lain.

Bayangkan sebuah peta kota raksasa di mana setiap kata memiliki alamat koordinat GPS.

- Kata "Raja" dan "Ratu" akan tinggal di lingkungan yang sama (lingkungan "Kerajaan").
- Kata "Kucing" dan "Anjing" akan bertetangga di distrik "Hewan Peliharaan".
- Jarak antara "Raja" dan "Mobil" akan sangat jauh karena mereka tidak memiliki hubungan semantik.

Kekuatan utama dari representasi ini adalah kemampuan untuk melakukan operasi matematika pada konsep abstrak, yang dikenal sebagai penalaran analogi.

2. Analogi Vektor dan Aljabar Semantik

Salah satu fenomena paling menarik dari word embeddings adalah kemampuan untuk menyelesaikan persamaan analogi melalui operasi aritmatika vektor. Contoh klasik yang sering dikutip adalah:

$$W_{raja} - W_{pria} + W_{wanita} \approx W_{ratu}$$

Penjelasannya, Jika Anda mengambil konsep "Raja", lalu membuang sifat maskulinitasnya ("Pria"), dan menambahkan sifat femininitas ("Wanita"), Anda akan mendarat di koordinat yang sangat dekat dengan "Ratu" dalam ruang vektor (Carl Alen, 2019).

C. DEEP LEARNING DENGAN PEMROSESAN SEKUENSIAL

Bahasa adalah data sekuensial; urutan kata menentukan makna. "Ibu memasak nasi" berbeda maknanya dengan "Nasi memasak ibu" (yang tidak masuk akal), meskipun kata-katanya sama. Arsitektur jaringan saraf standar (Feedforward Neural Networks) tidak memiliki memori tentang urutan ini, sehingga dikembangkanlah Recurrent Neural Networks (RNN).

1. Recurrent Neural Networks (RNN)

RNN dirancang untuk memproses urutan dengan cara mempertahankan "memori" dari langkah sebelumnya. Output dari pemrosesan kata pertama dimasukkan kembali sebagai input tambahan saat memproses kata kedua, dan seterusnya. Ini menciptakan siklus umpan balik (loop) yang memungkinkan informasi mengalir seiring waktu (Hassaan Idrees, 2024; Pascanu et al., 2013).

Bayangkan RNN seperti barisan orang yang bermain pesan berantai (Telephone Game).

- Orang pertama menerima kata, memprosesnya, dan membisikkan hasilnya ke orang kedua.
- Orang kedua menggabungkan bisikan tersebut dengan kata baru yang dia pegang, lalu membisikkannya ke orang ketiga.

Proses ini berlanjut hingga orang terakhir harus menyimpulkan makna kalimat berdasarkan akumulasi bisikan tersebut.

Kelemahan fatal RNN adalah ketidakmampuannya mengingat informasi dalam urutan yang panjang. Dalam kalimat panjang, informasi dari kata awal sering kali "luntur" atau menghilang sebelum mencapai akhir pemrosesan. Secara matematis, ini terjadi karena gradien (sinyal kesalahan yang digunakan untuk belajar) menjadi sangat kecil saat dikalikan berulang kali selama proses backpropagation, sehingga model berhenti belajar dari data masa lalu (SuperDataScience Team, 2018).

Dalam analogi pesan berantai, jika barisan orang terlalu panjang (misalnya 50 orang), pesan dari orang pertama hampir pasti sudah terdistorsi

atau hilang sama sekali saat mencapai orang terakhir. RNN memiliki ingatan yang sangat pendek.

2. Long Short-Term Memory (LSTM)

Untuk mengatasi "kepikunan" RNN, Hochreiter dan Schmidhuber (1997) memperkenalkan LSTM. Arsitektur ini menambahkan struktur internal yang kompleks pada setiap unit pemrosesan untuk mengatur aliran informasi, memungkinkan model mengingat dependensi jangka Panjang (Hochreiter & Schmidhuber, 1997).

LSTM memiliki tiga "gerbang" (gates) utama yang mengontrol sel memori:

- Forget Gate (Gerbang Lupa): Memutuskan informasi apa dari masa lalu yang tidak lagi relevan dan harus dibuang.
- Input Gate (Gerbang Masuk): Memutuskan informasi baru apa yang penting untuk disimpan.
- Output Gate (Gerbang Keluar): Menentukan informasi apa yang akan diteruskan ke langkah berikutnya.

Jika RNN adalah pesan berantai yang rentan lupa, LSTM ibarat sistem ban berjalan (conveyor belt) di sebuah pabrik canggih.

- Ada jalur khusus (Cell State) yang memungkinkan paket informasi penting meluncur mulus dari awal hingga akhir pabrik tanpa gangguan.
- Di sepanjang jalur, ada robot-robot cerdas (Gerbang). Satu robot bertugas membuang paket sampah (Forget Gate), sementara robot lain menambahkan paket baru yang penting (Input Gate).

Dengan mekanisme ini, informasi penting di awal kalimat (misalnya kata "tidak" dalam "tidak suka") dapat dijaga tetap utuh hingga akhir kalimat untuk memastikan klasifikasi sentimen yang akurat.

3. Gated Recurrent Unit (GRU)

GRU adalah varian yang lebih modern dan sederhana dari LSTM. GRU menggabungkan gerbang input dan forget menjadi satu Update Gate, serta menyatukan sel memori dan status tersembunyi (hidden state).

GRU memiliki parameter yang lebih sedikit daripada LSTM, membuatnya lebih cepat untuk dilatih dan membutuhkan daya komputasi yang lebih rendah, namun sering kali mencapai performa yang setara untuk banyak tugas NLP (Cho et al., 2014).

GRU adalah versi "efisiensi tinggi" dari mesin LSTM. Jika LSTM adalah mobil mewah dengan banyak tombol kontrol terpisah, GRU adalah mobil sport yang menyederhanakan kontrol tersebut menjadi lebih sedikit tombol namun tetap bertenaga.

D. CONVOLUTIONAL NEURAL NETWORKS

Meskipun CNN awalnya dirancang untuk visi komputer (Computer Vision), seperti mengenali wajah atau objek dalam gambar, arsitektur ini terbukti sangat efektif untuk klasifikasi teks (Conneau et al., 2017)

1. Mekanisme Konvulsi pada Data Teks

Dalam pengolahan citra, CNN menggeser filter di atas piksel gambar untuk menemukan pola visual (tepi, sudut, bentuk). Dalam NLP, CNN menggeser filter di atas urutan kata (matriks embedding) untuk menemukan pola linguistik lokal atau n-gram (kelompok kata yang berurutan) (Soni et al., 2022).

Bayangkan Anda memiliki kaca pembesar yang hanya bisa melihat tiga kata sekaligus. Anda menggeser kaca pembesar ini dari awal hingga akhir kalimat. Anda tidak mencoba memahami alur cerita keseluruhan secara mendalam (seperti RNN), tetapi Anda mencari "pola visual" tertentu.

- Jika kaca pembesar menemukan frasa "sangat buruk", sensor negatif aktif.
- Jika menemukan "luar biasa", sensor positif aktif.

CNN sangat efisien dalam mendeteksi frasa kunci yang menjadi indikator kuat sentimen, terlepas dari di mana frasa tersebut muncul dalam kalimat.

2. Mengambil Sinyal Terkuat (Pooling)

Setelah proses konvolusi, CNN menggunakan lapisan Pooling (biasanya Max-Pooling). Fungsinya adalah untuk mengambil nilai tertinggi dari hasil pemindaian filter (Jacovi et al., 2018).

Bayangkan sensor Anda berteriak dengan volume berbeda saat memindai kalimat. Max-Pooling hanya mencatat teriakan paling keras ("SANGAT BURUK!") dan mengabaikan gumaman latar belakang. Ini membuat model fokus pada fitur yang paling dominan (sangat positif atau sangat negatif) dan mengabaikan kata-kata pengisi yang tidak relevan.

Keuntungan utama CNN dibanding RNN adalah kecepatan. Karena operasi konvolusi dapat dilakukan secara paralel (semua bagian kalimat dipindai bersamaan), pelatihan CNN jauh lebih cepat daripada RNN yang harus memproses kata demi kata secara berurutan (Dertat, 2017).

E. TRANSFORMER

Pada tahun 2017, penelitian berjudul "Attention Is All You Need" oleh Vaswani et al. mengubah lanskap NLP selamanya dengan memperkenalkan arsitektur Transformer. Transformer meninggalkan rekurensi (proses berurutan)

sepenuhnya dan mengandalkan mekanisme yang disebut Self-Attention (Vaswani et al., 2017).

1. Mekanisme Self-Attention

Kelemahan RNN adalah harus memproses kata secara urut. Transformer memproses seluruh kalimat sekaligus (paralel). Untuk memahami konteks tanpa urutan waktu, Transformer menggunakan Self-Attention untuk menghitung hubungan antara setiap kata dengan semua kata lain dalam kalimat.

Bayangkan Anda berada di sebuah pesta yang ramai (kalimat panjang). Banyak orang berbicara sekaligus. Namun, telinga manusia memiliki kemampuan luar biasa untuk "fokus" pada satu suara tertentu yang relevan (misalnya seseorang menyebut nama Anda) dan meredam kebisingan lainnya

Dalam Transformer, setiap kata bertindak seperti orang di pesta ini. Kata tersebut "mendengarkan" semua kata lain dan memberikan perhatian lebih pada kata-kata yang relevan untuk memberikan konteks pada dirinya sendiri.

- Contoh: "Bank itu memberikan bunga tinggi."
- Kata "bunga" memiliki makna ganda (tanaman atau uang).
- Melalui self-attention, kata "bunga" akan memberikan bobot perhatian tinggi pada kata "Bank" dan "tinggi". Koneksi ini memberi tahu model bahwa dalam konteks ini, "bunga" bermakna finansial, bukan botani (Mehta, 2026)

2. Query, Key, dan Value

Secara teknis, mekanisme ini bekerja menggunakan tiga vektor untuk setiap kata: Query (Pertanyaan), Key (Kunci), dan Value (Nilai).

Analoginya kalua di Perpustakaan adalah berikut:

- Query (Q): Apa yang sedang dicari oleh sebuah kata? (Misal: Kata "dia" mencari referensi subjeknya).
- Key (K): Label pada buku-buku di rak (kata-kata lain dalam kalimat).
- Value (V): Isi informasi dari buku tersebut.

Model menghitung kecocokan antara Query satu kata dengan Key kata lain. Jika cocok, model mengambil Value (informasi) dari kata tersebut untuk memperkaya pemahaman (Mehta, 2026).

3. Bidirectional Encoder Representations from Transformer (BERT)

Varian Transformer yang paling berpengaruh dalam analisis sentimen adalah BERT (Devlin et al., 2019). Inovasi utama BERT adalah sifatnya yang Bidirectional (dua arah.Model sebelumnya (seperti GPT awal atau RNN) membaca teks secara searah (kiri ke kanan). BERT membaca seluruh kalimat

Seperti membaca novel misteri. Untuk benar-benar memahami petunjuk di halaman 10, Anda sering kali perlu mengetahui apa yang terjadi di halaman 1 dan apa yang terungkap di halaman 50. BERT melihat "seluruh buku" secara simultan untuk memahami konteks penuh dari setiap kata, membuatnya sangat akurat dalam menangkap ambiguitas Bahasa (Talebi Shaw, 2024).

F. TRANSFORMER DALAM KONTEKS BAHASA INDONESIA

Meskipun model seperti BERT sangat kuat, model yang dilatih hanya dengan data Bahasa Inggris sering gagal menangkap nuansa Bahasa Indonesia, apalagi bahasa daerah. Bahasa Indonesia memiliki morfologi aglutinatif (kaya imbuhan) dan struktur kalimat fleksibel yang berbeda dari Bahasa Inggris. Oleh karena itu, pengembangan model monolingual (satu bahasa) menjadi krusial.

1. IndoBERT dan IndoLEM

IndoBERT adalah model berbasis BERT yang dilatih khusus menggunakan korpus teks Bahasa Indonesia yang massif (Rihaadatul'Aisy & Pamungkas, 2025)

- Data Pelatihan: IndoBERT dilatih pada sekitar 4 miliar kata yang diambil dari Wikipedia Bahasa Indonesia, artikel berita, dan data web lainnya (Rihaadatul'Aisy & Pamungkas, 2025).
- Performa Komparatif:
- Dalam tugas analisis sentimen pada ulasan hotel, IndoBERT mencapai akurasi 92.52%, dengan presisi 1.00 untuk kelas negative (Singgalen, 2025).
- Pada dataset berita, IndoBERT mencatat akurasi 70.76% (F1-score 0.71), mengungguli model IndoRoBERTa (68.79%) dan jauh di atas baseline SVM yang hanya 61% (Rihaadatul'Aisy & Pamungkas, 2025).

Studi lain pada ulasan aplikasi kesehatan menunjukkan IndoBERT mencapai akurasi hingga 96% (Imaduddin et al., n.d.)

Data ini menegaskan bahwa model yang "lahir dan besar" dengan data lokal memiliki intuisi bahasa yang jauh lebih tajam daripada model multilingual umum.

2. Model untuk Bahasa Daerah (Jawa dan Sunda)

Indonesia memiliki ratusan bahasa daerah. Model analisis sentimen nasional sering kali gagal memahami ulasan yang menggunakan bahasa daerah atau campuran (code-mixing). Inisiatif seperti NusaX (Winata et al., 2023)dan model dari komunitas Hugging Face (w11wo) berusaha mengisi kekosongan ini.

Tabel Perbandingan Model Bahasa Daerah dan Performanya (w11wo, n.d.-a, n.d.-c, n.d.-b; Wongso et al., 2021)

Model	Arsitektur Dasar	Dataset Pelatihan	Metrik Performa	Keterangan
Javanese BERT Small (w11wo)	BERT-Small (110M Params)	Wikipedia Jawa (319 MB)	Perplexity: 22.00	Model dasar *Masked Language Model*.
Javanese BERT IMDB Classifier	BERT-Small	Terjemahan IMDB (Jawa)	Akurasi: 76.37%	Fine-tuned untuk sentimen film.
Sundanese RoBERTa Base (w11wo)	RoBERTa-Base (124M Params)	OSCAR, mC4, Wiki Sunda (758 MB)	Akurasi MLM: 63.98%	Dilatih *from scratch* dengan Flax.
IndoRoBERTa (w11wo)	RoBERTa-Base	SmSA (IndoNLU)	Akurasi: 94.36%	Sangat efektif untuk bahasa informal/medsos.

Perbedaan akurasi antara model Bahasa Indonesia (~94%) dan Bahasa Jawa (~76%) menunjukkan tantangan low-resource language. Data pelatihan untuk bahasa daerah jauh lebih sedikit, sehingga model kurang terpapar pada variasi kosakata dan struktur kalimat. Meskipun demikian, keberadaan model ini memungkinkan deteksi sentimen pada kalimat seperti "Pelayanane ora mbejaji" (Jawa: Pelayanannya tidak berharga/buruk) yang akan terlewatkan oleh model standar.

G. FINE-TUNING

Memiliki model pre-trained seperti IndoBERT ibarat memiliki lulusan sastra yang cerdas. Namun, untuk menjadikannya ahli analisis sentimen ulasan produk, ia perlu dilatih ulang secara spesifik (fine-tuning).

1. Konsep Transfer-Learning

Transfer Learning adalah metode di mana pengetahuan yang dipelajari dari satu tugas (misalnya memahami bahasa secara umum) diterapkan pada tugas lain yang terkait (klasifikasi sentimen).

Kalau kita Analogikan dengan Dokter Spesialis, maka

- Pre-training: Seperti sekolah kedokteran umum. Mahasiswa belajar anatomi dan fisiologi dasar (struktur bahasa). Proses ini mahal dan lama (ribuan jam GPU).
- Fine-tuning: Seperti masa residensi spesialis. Dokter umum belajar keahlian khusus (jantung, mata). Proses ini cepat dan hanya butuh sedikit data spesifik (Bergmann, 2026).

2. Langkah Fine-Tuning

a. Tokenisasi (Tokenization)

Mengubah teks menjadi token angka. BERT menggunakan WordPiece atau BPE.

- Penting untuk Bahasa Indonesia: Tokenizer memecah kata kompleks menjadi sub-kata. Kata "memperjuangkan" mungkin dipecah menjadi mem, ##per, ##juang, ##kan. Ini memungkinkan model memahami akar kata "juang" meskipun ditempeli berbagai imbuhan (Poomka et al., 2021).
- Padding & Truncation: Menyamakan panjang semua kalimat input (misal 128 token). Kalimat pendek ditambah angka 0 (pad), kalimat panjang dipotong.

b. Konfigurasi Hyperparameter

- Learning Rate (Laju Pembelajaran): Harus sangat kecil (misal 2e-5 hingga 5e-5). Karena model sudah "pintar", kita hanya perlu melakukan penyesuaian bobot yang halus. Jika terlalu besar, pengetahuan lama model bisa rusak (Catastrophic Forgetting).
- Epochs: Biasanya 2-4 epoch sudah cukup. Terlalu lama bisa menyebabkan overfitting.
- Batch Size: Tergantung memori GPU, biasanya 16 atau 32.(Eltahier et al., 2025)

c. Regulasi (Dropout)

Teknik mematikan sebagian neuron secara acak selama pelatihan untuk mencegah model terlalu bergantung pada pola tertentu (menghafal data latih).

d. Optimizer (AdamW)

Algoritma optimasi yang populer untuk Transformer. AdamW (Adam dengan Weight Decay) membantu model mencapai konvergensi lebih stabil (Loshchilov & Hutter, 2017).

H. TANTANGAN IMPLEMENTASI DEEP LEARNING

Meskipun Deep Learning menawarkan performa superior, implementasinya di dunia nyata menghadapi sejumlah tantangan signifikan.

1. Tantangan Overfitting dan Biaya Komputasi

Model Transformer memiliki jumlah parameter yang masif (IndoBERT ~110 juta, GPT-3 ~175 miliar).

- Overfitting pada Data Kecil: Model raksasa cenderung "menghafal" data jika dataset pelatihan terlalu kecil (seperti pada kasus bahasa daerah). Solusinya adalah penggunaan teknik regularisasi agresif atau Data Augmentation (Overfitting In Sentiment Analysis, 2025).
- Biaya Sumber Daya: Melatih dan menjalankan (inference) model ini membutuhkan GPU yang mahal dan konsumsi energi tinggi. Ini menjadi hambatan bagi UMKM atau peneliti independen. Solusi masa depan mengarah pada Teknik distillation (membuat model guru besar mengajar model murid yang kecil dan cepat) (Deepak, 2024).

2. Tantangan Interpretabilitas

Salah satu kelemahan terbesar Deep Learning dibanding ML tradisional adalah sifatnya sebagai "Kotak Hitam" (Black Box). Sulit untuk menjelaskan mengapa model memutuskan sebuah kalimat bersentimen negatif.

- Pada ML tradisional, kita bisa melihat bobot fitur ("kata 'jelek' punya skor -5").

- Pada DL, keputusan tersebar di jutaan neuron. Ini menjadi masalah etika dan kepercayaan, terutama di sektor sensitif seperti keuangan atau kesehatan. Teknik Attention Visualization mulai digunakan untuk memetakan fokus model, namun ini masih belum sempurna (Islam et al., 2024).

3. Masa Depan Multimodal LLM

Tren menuju tahun 2025 menunjukkan pergeseran dari analisis teks murni menuju Multimodal Large Language Models (LLMs). Analisis sentimen masa depan tidak hanya akan membaca teks ulasan, tetapi juga:

- Menganalisis nada suara (audio) dari keluhan pelanggan.
- Membaca ekspresi mikro wajah (visual) dari video testimoni.
- Menggabungkan semua modalitas ini untuk mendapatkan pemahaman emosi yang holistik dan akurat (Bill, 2025).

Selain itu, model generatif (seperti GPT-4 atau Llama) mulai menggantikan model klasifikasi khusus. Alih-alih melatih model khusus sentimen, kita bisa menggunakan Prompt Engineering pada LLM: "Analisis sentimen teks berikut dan jelaskan alasannya." Pendekatan ini menawarkan fleksibilitas dan penjelasan (explainability) yang lebih baik (TECHSUR, 2025).

Transisi dari Machine Learning ke Deep Learning dalam analisis sentimen bukan sekadar peningkatan akurasi, melainkan perubahan fundamental dalam cara mesin memahami bahasa manusia. Kita telah bergerak dari pencocokan kata kunci yang kaku (kamus), menuju pemahaman pola statistik (ML), dan akhirnya tiba pada pemahaman kontekstual mendalam melalui representasi vektor dan mekanisme atensi (Transformer).

Bagi ekosistem teknologi Indonesia, ketersediaan model pre-trained lokal seperti IndoBERT dan inisiatif bahasa daerah adalah aset strategis yang tak ternilai. Meskipun tantangan teknis seperti kebutuhan komputasi dan data low-resource masih ada, teknik fine-tuning yang tepat memungkinkan adaptasi model canggih ini untuk berbagai aplikasi industri. Dengan memahami arsitektur di balik model-model ini, dari gerbang LSTM hingga atensi Transformer, para praktisi dapat membangun sistem analisis sentimen yang tidak hanya cerdas, tetapi juga peka terhadap nuansa budaya dan bahasa lokal.

BAB 9. EVALUASI DAN VALIDASI MODEL ANALISIS SENTIMEN

Ayun Hapsari, S.Kom.

A. KONSEP DASAR EVALUASI MODEL

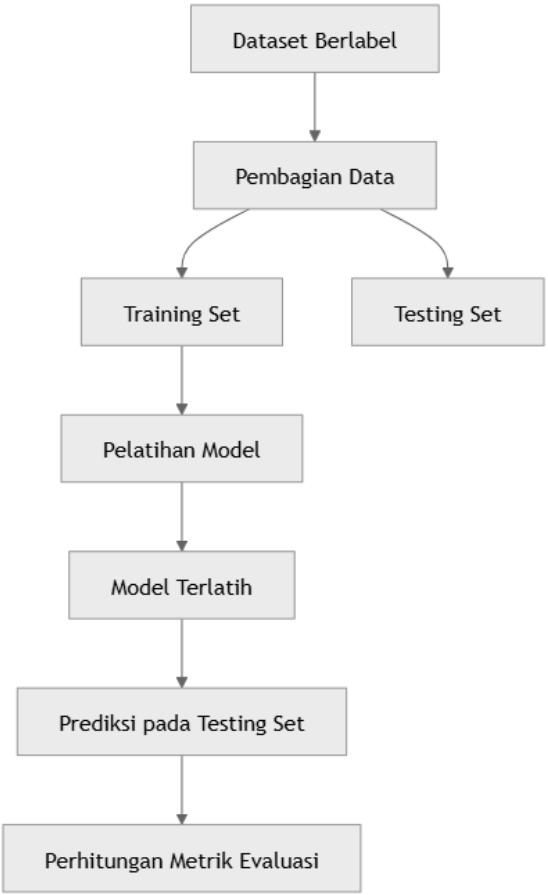
Evaluasi model bertujuan mengukur seberapa baik sistem melakukan prediksi pada data baru. Dalam analisis sentimen, evaluasi menentukan apakah model mampu mengklasifikasikan opini secara akurat. Tanpa evaluasi yang tepat, hasil pelatihan tidak memiliki makna praktis. Proses ini membantu memastikan bahwa model tidak hanya bekerja pada data latih.

Tujuan utama evaluasi dalam machine learning adalah mengukur kemampuan generalisasi. Generalisasi merujuk pada kemampuan model memprediksi data yang belum pernah dilihat sebelumnya. Model yang baik tidak hanya menghafal data latih. Model harus mampu mengenali pola yang konsisten pada data baru.

Perbedaan antara evaluasi pelatihan dan evaluasi pengujian sangat penting. Evaluasi pelatihan dilakukan pada data yang digunakan untuk membangun model. Nilai performa pada tahap ini sering lebih tinggi karena model telah melihat data tersebut. Evaluasi pengujian dilakukan pada data terpisah yang tidak digunakan selama pelatihan. Nilai ini mencerminkan performa yang lebih realistis.

Hubungan antara generalisasi dan overfitting perlu diperhatikan. Overfitting terjadi ketika model terlalu menyesuaikan diri pada data latih. Model dapat mencapai akurasi tinggi pada training set tetapi gagal pada testing set. Kondisi ini menunjukkan bahwa model menangkap noise alih-alih pola umum. Evaluasi pada data terpisah membantu mendeteksi masalah ini.

Alur umum evaluasi model dapat digambarkan sebagai berikut.



Alur Evaluasi Model

Evaluasi juga berperan dalam pengambilan keputusan penelitian. Peneliti menggunakan hasil evaluasi untuk memilih model terbaik. Jika dua model memiliki akurasi berbeda, evaluasi membantu menentukan pilihan yang lebih stabil. Selain itu, hasil evaluasi digunakan untuk membandingkan metode dalam studi akademik.

Evaluasi bukan sekadar perhitungan angka. Proses ini menentukan validitas kesimpulan penelitian. Model dengan performa tinggi pada data uji memberikan dasar yang lebih kuat untuk interpretasi. Oleh karena itu, pemahaman konsep dasar evaluasi menjadi fondasi sebelum membahas metrik klasifikasi secara lebih rinci.

B. CONFUSION MATRIX

Confusion matrix adalah tabel yang merangkum hasil prediksi model dibandingkan dengan label sebenarnya. Matriks ini membantu memahami jenis kesalahan yang dibuat model. Dalam klasifikasi biner sentimen, kelas biasanya terdiri dari positif dan negatif.

Struktur confusion matrix biner dapat ditampilkan sebagai berikut.

	Prediksi Positif	**Prediksi Negatif**
Aktual Positif	True Positive (TP)	False Negative (FN)
Aktual Negatif	False Positive (FP)	True Negative (TN)

True Positive terjadi ketika model memprediksi positif dan labelnya memang positif. True Negative terjadi ketika model memprediksi negatif dan labelnya memang negatif. False Positive terjadi ketika model memprediksi positif tetapi label sebenarnya negatif. False Negative terjadi ketika model memprediksi negatif padahal sebenarnya positif.

Interpretasi kesalahan sangat penting dalam analisis sentimen. False Positive berarti sistem salah menilai opini negatif sebagai positif. False Negative berarti opini positif gagal dikenali. Dalam konteks evaluasi produk, kedua jenis kesalahan memiliki implikasi berbeda.

C. ACCURACY, PRECISION, RECALL, DAN F1-SCORE

Empat metrik ini merupakan ukuran utama dalam evaluasi model klasifikasi sentimen. Masing-masing metrik menilai performa dari sudut yang berbeda. Pemahaman yang tepat membantu peneliti memilih ukuran yang sesuai dengan tujuan analisis.

Accuracy mengukur proporsi prediksi yang benar terhadap seluruh data. Rumusnya adalah:

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

Accuracy mudah dipahami dan sering digunakan sebagai indikator awal. Namun metrik ini dapat menyesatkan pada data tidak seimbang. Jika sebagian besar data termasuk satu kelas, model dapat mencapai accuracy tinggi tanpa benar-benar memahami pola.

Precision mengukur ketepatan prediksi positif. Rumusnya adalah:

$$Precision = \frac{TP}{TP+FP}$$

Precision tinggi menunjukkan bahwa sebagian besar prediksi positif memang benar. Metrik ini penting ketika kesalahan false positive harus diminimalkan. Dalam analisis reputasi merek, kesalahan menganggap opini negatif sebagai positif dapat berdampak besar.

Recall mengukur kemampuan model mendeteksi seluruh contoh positif. Rumusnya adalah:

$$Recall = \frac{TP}{TP+FN}$$

Recall tinggi berarti model jarang melewatkan kasus positif. Metrik ini relevan ketika false negative memiliki konsekuensi serius. Dalam analisis keluhan pelanggan, kegagalan mendeteksi opini negatif dapat merugikan organisasi.

F1-score menggabungkan precision dan recall dalam satu ukuran. Rumusnya adalah:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score menggunakan rata-rata harmonik sehingga menyeimbangkan kedua metrik. Nilai tinggi menunjukkan bahwa model memiliki ketepatan dan kelengkapan deteksi yang baik. Pada dataset tidak seimbang, F1-score sering lebih informatif dibanding accuracy.

Perbandingan singkat keempat metrik dapat dilihat pada tabel berikut.

Metrik	Fokus Utama	Kelebihan	Keterbatasan
Accuracy	Prediksi benar keseluruhan	Mudah dipahami	Bias pada data tidak seimbang
Precision	Ketepatan kelas positif	Kurangi false positive	Abaikan false negative
Recall	Kelengkapan deteksi positif	Kurangi false negative	Abaikan false positive
F1-score	Keseimbangan precision dan recall	Stabil pada data tidak seimbang	Tidak mempertimbangkan TN

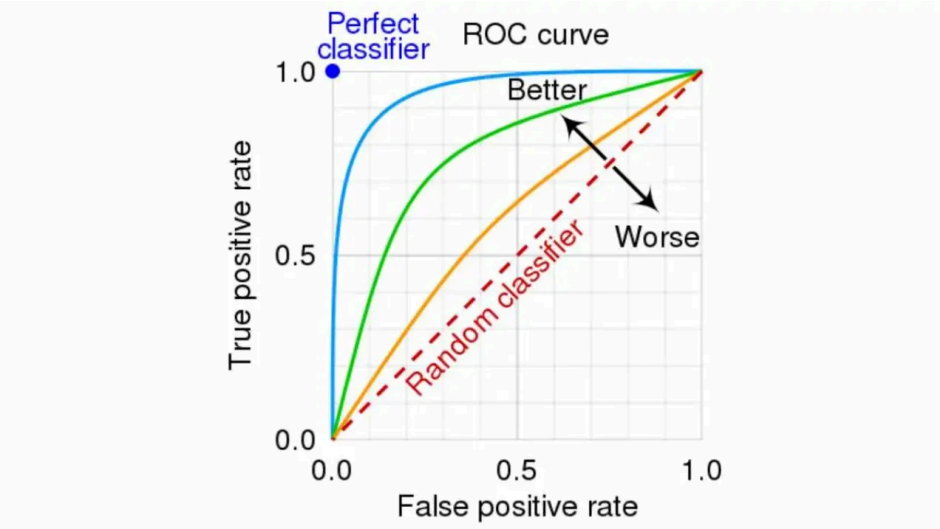
Pemilihan metrik harus disesuaikan dengan konteks penelitian. Jika distribusi kelas seimbang, accuracy dapat digunakan. Jika data tidak seimbang, F1-score atau kombinasi precision dan recall lebih tepat. Evaluasi yang komprehensif biasanya menggunakan lebih dari satu metrik agar hasil lebih akurat dan tidak menyesatkan.

D. ROC CURVE DAN AUC

ROC Curve menggambarkan hubungan antara True Positive Rate dan False Positive Rate pada berbagai ambang keputusan. True Positive Rate identik dengan recall. False Positive Rate dihitung sebagai:

$$FPR = \frac{FP}{FP+TN}$$

Kurva ROC menunjukkan trade-off antara sensitivitas dan kesalahan positif. Model yang baik memiliki kurva yang mendekati sudut kiri atas.



Ilustrasi Kurva ROC

AUC atau Area Under Curve mengukur luas di bawah kurva ROC. Nilai AUC berkisar antara 0 dan 1. Nilai mendekati 1 menunjukkan performa sangat baik. AUC memudahkan perbandingan model tanpa bergantung pada satu ambang keputusan.

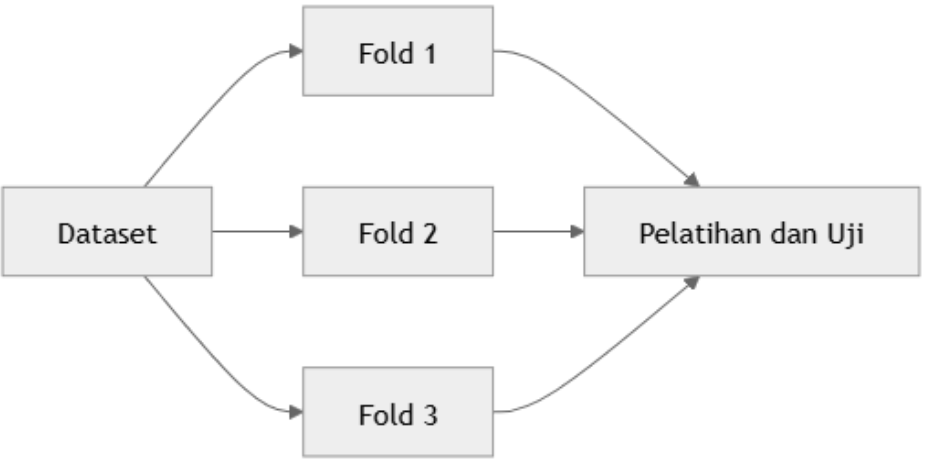
E. VALIDASI MODEL

Train-test split adalah metode sederhana yang membagi data menjadi dua bagian. Biasanya 70 hingga 80 persen digunakan untuk pelatihan, sisanya untuk pengujian. Metode ini mudah diterapkan tetapi sensitif terhadap pembagian acak.

Cross validation membagi data menjadi beberapa bagian dan melakukan pelatihan berulang. K-fold cross validation membagi data menjadi K bagian. Model dilatih sebanyak K kali dengan satu bagian sebagai data uji setiap kali.

Stratified K-fold memastikan distribusi kelas tetap seimbang di setiap fold. Metode ini penting pada dataset tidak seimbang.

Alur K-fold cross validation dapat digambarkan sebagai berikut.



Cara Kerja Cross Validation

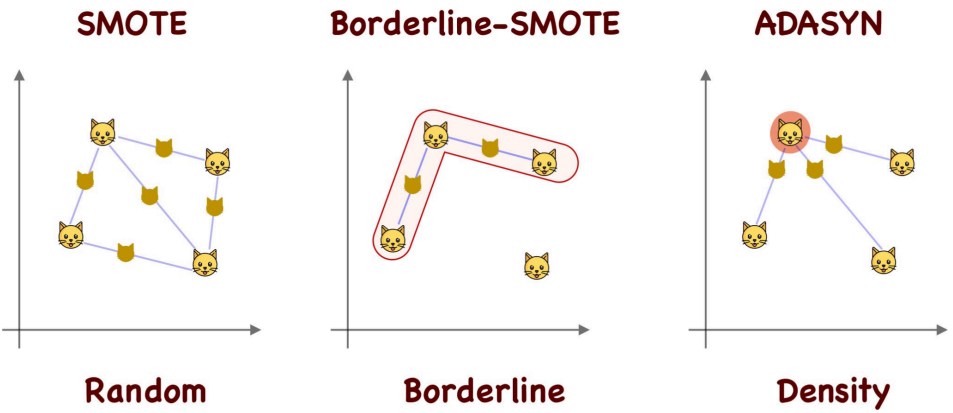
Cross validation memberikan estimasi performa yang lebih stabil dibanding satu kali split.

F. IMBALANCED DATA DALAM ANALISIS SENTIMEN

Data tidak seimbang terjadi ketika satu kelas jauh lebih dominan. Dalam analisis sentimen, opini positif sering lebih banyak daripada negatif. Ketidakseimbangan ini dapat menyesatkan metrik accuracy.

Beberapa teknik digunakan untuk menangani kondisi ini. Oversampling menambah jumlah data pada kelas minoritas. Undersampling mengurangi data pada kelas mayoritas. SMOTE menghasilkan contoh sintetis untuk kelas minoritas. Class weighting memberikan penalti lebih besar pada kesalahan kelas

minoritas. Pemilihan metrik juga harus disesuaikan. F1-score dan AUC lebih informatif dibanding accuracy pada kondisi ini.



Penanganan Imbalanced Data

SMOTE, Borderline-SMOTE, dan ADASYN merupakan tiga teknik utama untuk menangani data tidak seimbang melalui pembuatan sampel sintetis, namun ketiganya memiliki strategi penempatan data yang berbeda.

SMOTE bekerja secara merata dengan menciptakan sampel baru di sepanjang garis lurus yang menghubungkan titik data minoritas dengan tetangga terdekatnya tanpa mempedulikan posisi data tersebut terhadap kelas mayoritas.

Sementara itu, Borderline-SMOTE melakukan pendekatan yang lebih spesifik dengan hanya menghasilkan data sintetis di area perbatasan atau zona bahaya di mana sampel minoritas dikelilingi oleh banyak sampel mayoritas, sehingga memperkuat batas keputusan model.

Di sisi lain, ADASYN bertindak lebih dinamis dengan memberikan bobot lebih pada sampel minoritas yang paling sulit dipelajari; algoritma ini secara adaptif menciptakan lebih banyak data baru di area dengan kepadatan rendah untuk membantu model mengatasi kesulitan klasifikasi di wilayah tersebut.

G. ERROR ANALYSIS

Error analysis membantu memahami kesalahan model secara mendalam. Proses ini melibatkan pemeriksaan manual terhadap prediksi yang salah. Peneliti dapat mengidentifikasi pola kesalahan tertentu.

Analisis terhadap false positive dan false negative sering mengungkap masalah pra-pemrosesan atau representasi fitur. Sarkasme dan ambiguitas kata sering menjadi sumber kesalahan. Dengan memahami pola ini, peneliti dapat memperbaiki model.

Perbaikan dapat dilakukan melalui penyesuaian fitur, penggantian algoritma, atau penambahan data latih. Dokumentasi hasil evaluasi dan analisis kesalahan penting dalam penelitian akademik. Transparansi ini memperkuat validitas dan kredibilitas hasil penelitian.

BAB 10. IMPLEMENTASI SISTEM ANALISIS SENTIMEN

Egi Dio Bagus Sudewo, S.Kom., M.Kom.

A. KONSEP IMPLEMENTASI SISTEM ANALISIS SENTIMEN

Model yang dikembangkan dalam penelitian sering berbeda dengan sistem yang digunakan dalam produksi. Pada tahap eksperimen, fokus utama terletak pada akurasi dan evaluasi. Peneliti bekerja dengan dataset terkontrol dan lingkungan komputasi tertentu. Model dapat dijalankan secara lokal tanpa mempertimbangkan beban pengguna.

Sistem produksi memiliki tuntutan berbeda. Sistem harus melayani permintaan secara real-time atau hampir real-time. Model tidak hanya diuji pada data statis, tetapi juga pada data yang terus berubah. Stabilitas dan kecepatan menjadi faktor penting selain akurasi.

Perbedaan lain terletak pada pengelolaan data. Dalam eksperimen, data biasanya sudah dibersihkan dan diberi label. Dalam sistem nyata, data masuk secara dinamis dan belum tentu terstruktur. Sistem harus mampu menangani variasi input, kesalahan format, dan potensi spam.

Komponen utama sistem analisis sentimen terdiri dari beberapa bagian. Pertama, modul akuisisi data yang menerima input dari pengguna atau sumber eksternal. Kedua, modul pra-pemrosesan yang membersihkan dan menormalisasi teks. Ketiga, modul representasi fitur yang mengubah teks menjadi vektor numerik. Keempat, modul prediksi yang memuat model terlatih. Terakhir, modul penyimpanan dan visualisasi hasil.

Hubungan antar komponen tersebut dapat digambarkan sebagai berikut.

Komponen Analisis Sentimen

Tantangan implementasi di lingkungan nyata cukup kompleks. Sistem harus menangani lonjakan permintaan pengguna. Waktu respons yang lambat

dapat menurunkan kualitas layanan. Selain itu, model perlu diperbarui secara berkala agar tetap relevan dengan tren bahasa terbaru.

Keamanan juga menjadi perhatian penting. Sistem berbasis web harus memvalidasi input untuk mencegah penyalahgunaan. Data pengguna perlu dikelola dengan hati-hati sesuai regulasi perlindungan data. Integrasi dengan database dan API eksternal juga memerlukan pengamanan yang memadai.

Kriteria sistem yang baik mencakup beberapa aspek utama. Sistem harus akurat dalam memprediksi sentimen. Sistem harus stabil dan tidak mudah gagal ketika menerima banyak permintaan. Sistem juga harus responsif dengan waktu prediksi yang singkat. Kombinasi ketiga aspek tersebut menentukan keberhasilan implementasi.

Implementasi bukan sekadar memindahkan model ke server. Proses ini melibatkan desain arsitektur, pengelolaan sumber daya, dan pemantauan performa. Dengan pemahaman konsep dasar ini, pembahasan dapat dilanjutkan pada pipeline sistem secara lebih rinci.

B. PIPELINE SISTEM ANALISIS SENTIMEN

Pipeline sistem menggambarkan alur kerja dari input hingga output akhir. Struktur ini membantu memastikan setiap tahap berjalan sistematis. Pipeline yang jelas juga memudahkan pemeliharaan dan pengembangan sistem. Setiap komponen memiliki peran yang saling terhubung.

Tahap pertama adalah akuisisi data. Sistem menerima teks dari pengguna, API eksternal, atau database. Input dapat berupa satu kalimat atau kumpulan dokumen. Pada tahap ini, sistem juga melakukan validasi format. Validasi penting untuk mencegah kesalahan proses berikutnya.

Tahap kedua adalah pra-pemrosesan. Teks dibersihkan dari URL, simbol berlebih, dan karakter tidak relevan. Sistem melakukan tokenisasi dan normalisasi sesuai kebutuhan model. Jika model menggunakan embedding tertentu, format input harus disesuaikan. Tahap ini memastikan konsistensi representasi.

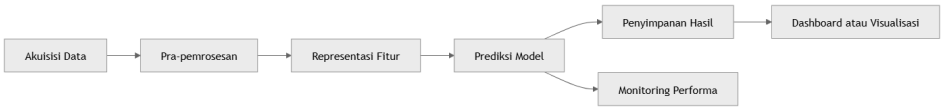
Tahap ketiga adalah representasi fitur. Teks diubah menjadi vektor numerik. Jika sistem menggunakan TF-IDF, maka teks dipetakan ke matriks fitur. Jika menggunakan embedding, teks diproses menjadi representasi dense. Konsistensi antara pelatihan dan produksi harus dijaga pada tahap ini.

Tahap keempat adalah prediksi model. Model terlatih dimuat ke dalam memori server. Sistem menghitung probabilitas atau skor sentimen berdasarkan input. Hasil prediksi biasanya berupa label dan nilai probabilitas. Waktu respons pada tahap ini harus dijaga tetap singkat.

Tahap kelima adalah penyimpanan dan visualisasi hasil. Hasil prediksi dapat disimpan dalam database untuk analisis lanjutan. Sistem juga dapat menampilkan hasil dalam bentuk dashboard. Visualisasi membantu pengguna memahami distribusi sentimen secara cepat.

Tahap terakhir adalah monitoring performa sistem. Monitoring mencakup waktu respons, tingkat kesalahan, dan beban server. Jika performa menurun, sistem dapat diperbarui atau dioptimalkan. Monitoring juga membantu mendeteksi perubahan pola bahasa yang memengaruhi akurasi.

Alur pipeline sistem dapat digambarkan sebagai berikut.



Alur Pipeline Sistem

Pipeline yang terstruktur membantu menjaga konsistensi sistem. Setiap tahap dapat diperbaiki tanpa mengubah keseluruhan arsitektur. Pendekatan modular ini mendukung pengembangan jangka panjang dan integrasi dengan layanan lain. Dengan pipeline yang jelas, implementasi sistem menjadi lebih stabil dan terukur.

C. ARSITEKTUR SISTEM ANALISIS SENTIMEN

Arsitektur sistem menentukan bagaimana komponen perangkat lunak saling terhubung. Desain arsitektur memengaruhi skalabilitas, keamanan, dan performa sistem. Dalam implementasi analisis sentimen, arsitektur dapat dirancang sederhana atau lebih terdistribusi. Pilihan bergantung pada kebutuhan pengguna dan beban sistem.

Arsitektur sederhana berbasis satu server cocok untuk prototipe atau skala kecil. Pada pendekatan ini, seluruh komponen berada dalam satu mesin. Server menangani permintaan pengguna, memproses teks, menjalankan model, dan menyimpan hasil. Struktur ini mudah dikembangkan dan dipelihara pada tahap awal. Namun kapasitasnya terbatas jika jumlah pengguna meningkat.

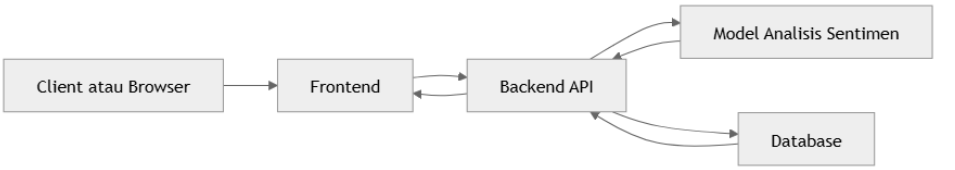
Pada arsitektur client–server, sistem dibagi menjadi dua bagian utama. Client berfungsi sebagai antarmuka pengguna, biasanya berupa aplikasi web atau mobile. Server menangani logika bisnis dan pemrosesan model. Pembagian

ini meningkatkan fleksibilitas dan memungkinkan pengelolaan beban lebih baik. Sistem juga lebih mudah dikembangkan secara modular.

Integrasi database menjadi komponen penting dalam sistem produksi. Database menyimpan input teks, hasil prediksi, dan log aktivitas. Penyimpanan ini memungkinkan analisis historis dan pelaporan. Sistem dapat menggunakan database relasional atau non-relasional sesuai kebutuhan. Integrasi yang baik membantu menjaga konsistensi data.

Komponen backend bertanggung jawab atas pemrosesan inti. Backend mengelola pra-pemrosesan, representasi fitur, dan pemanggilan model. Backend juga menyediakan API untuk menerima dan mengirim data. Komponen frontend berfungsi menampilkan hasil kepada pengguna. Frontend dapat menampilkan grafik distribusi sentimen dan tren waktu.

Diagram arsitektur umum sistem dapat digambarkan sebagai berikut.



Arsitektur Sistem Analisis Sentimen

Pada diagram tersebut, client mengirim permintaan melalui frontend. Backend menerima permintaan dan memanggil model untuk melakukan prediksi. Hasil prediksi disimpan dalam database dan dikirim kembali ke frontend. Struktur ini memungkinkan pemisahan tanggung jawab antar komponen.

Pemilihan arsitektur harus mempertimbangkan kebutuhan skalabilitas dan keamanan. Sistem dengan jumlah pengguna besar memerlukan server yang dapat ditingkatkan kapasitasnya. Penggunaan container atau layanan cloud dapat menjadi solusi pada tahap lanjutan. Dengan arsitektur yang tepat, sistem analisis sentimen dapat berjalan stabil dan responsif dalam jangka panjang.

D. DEPLOYMENT MODEL

Deployment model adalah proses mengubah model terlatih menjadi layanan yang dapat diakses pengguna. Pada tahap eksperimen, model biasanya dijalankan di lingkungan lokal seperti notebook. Pada tahap produksi, model harus tersedia melalui server atau aplikasi web. Proses ini membutuhkan penyesuaian struktur dan manajemen dependensi.

Langkah pertama adalah konversi model menjadi layanan. Model disimpan dalam format tertentu setelah pelatihan selesai. Format umum meliputi file pickle, joblib, atau format khusus deep learning seperti .pt atau .h5. File ini kemudian dimuat kembali pada server saat sistem berjalan.

Penyimpanan model terlatih harus dilakukan secara terkontrol. Versi model perlu dicatat agar perubahan dapat dilacak. Praktik ini disebut model versioning. Jika performa model menurun, sistem dapat kembali ke versi sebelumnya. Dokumentasi konfigurasi pelatihan juga perlu disimpan bersama model.

Loading model pada server dilakukan saat aplikasi dijalankan. Model dimuat ke dalam memori agar proses prediksi berjalan cepat. Jika model terlalu besar, waktu inisialisasi dapat meningkat. Oleh karena itu, optimasi ukuran model menjadi penting pada tahap deployment.

Pengujian endpoint dilakukan setelah model tersedia sebagai layanan. Endpoint adalah alamat yang menerima permintaan prediksi. Sistem harus diuji dengan berbagai jenis input untuk memastikan respons stabil. Pengujian juga mencakup validasi format JSON dan penanganan kesalahan.

Alur umum deployment dapat digambarkan sebagai berikut.

Alur Deployment

1. Deployment menggunakan Flask

Flask adalah framework Python yang ringan untuk membangun web service. Model dimuat pada server Flask dan dihubungkan dengan endpoint tertentu. Pengguna mengirim teks melalui permintaan HTTP. Server memproses teks dan mengembalikan hasil dalam format JSON.

Logo Flask

Flask cocok untuk membangun REST API sederhana. Struktur aplikasinya fleksibel dan mudah dikembangkan. Namun pengelolaan antarmuka visual memerlukan integrasi tambahan dengan frontend terpisah.

2. Deployment menggunakan Streamlit

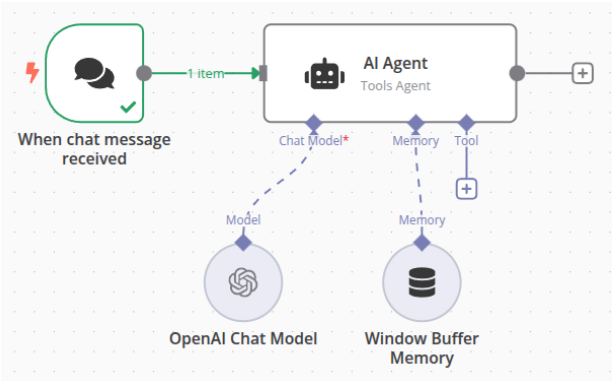
Streamlit dirancang untuk membangun aplikasi data interaktif dengan cepat. Model dapat dimuat dan langsung dihubungkan dengan komponen antarmuka seperti kotak teks dan grafik. Pendekatan ini cocok untuk demonstrasi dan prototipe.

Logo Streamlit

Keunggulan Streamlit terletak pada kemudahan penggunaan. Pengguna dapat melihat hasil prediksi secara langsung melalui dashboard. Namun kontrol terhadap arsitektur backend lebih terbatas dibanding Flask.

3. **Deployment dengan n8n**

n8n adalah platform otomasi berbasis workflow. Sistem dapat menghubungkan model analisis sentimen dengan layanan lain seperti email, database, atau media sosial. Model dapat dipanggil melalui webhook atau API.



n8n Workflow

Pendekatan ini cocok untuk integrasi otomatis dalam alur kerja bisnis. Misalnya, setiap komentar baru dapat langsung dianalisis dan dikategorikan. Namun n8n bergantung pada integrasi API yang sudah tersedia.

Perbandingan ketiga pendekatan dapat dilihat pada tabel berikut.

Aspek	Flask	Streamlit	n8n
Fokus utama	REST API	Dashboard interaktif	Otomasi workflow
Kemudahan setup	Sedang	Sangat mudah	Mudah
Kontrol backend	Tinggi	Terbatas	Bergantung integrasi
Cocok untuk	Sistem produksi API	Demo dan prototipe	Integrasi proses bisnis

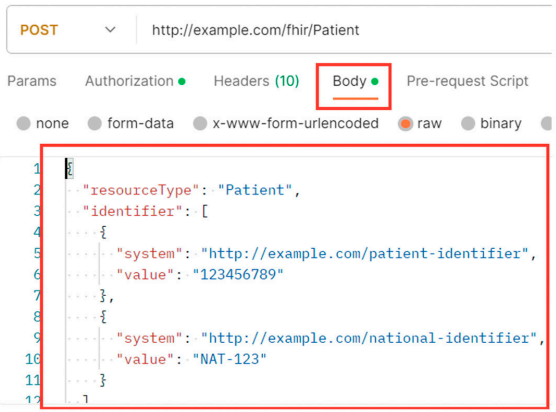
Pemilihan metode deployment bergantung pada tujuan sistem. Jika dibutuhkan API fleksibel, Flask menjadi pilihan tepat. Jika tujuan utama adalah visualisasi cepat, Streamlit lebih sesuai. Jika sistem perlu terintegrasi dengan banyak layanan, n8n memberikan solusi praktis. Dengan strategi deployment yang tepat, model analisis sentimen dapat diakses dan dimanfaatkan secara luas.

E. REST API UNTUK ANALISIS SENTIMEN

REST API adalah mekanisme komunikasi antara client dan server melalui protokol HTTP. REST singkatan dari Representational State Transfer. Konsep ini menggunakan metode standar seperti GET dan POST. Dalam sistem analisis sentimen, REST API memungkinkan pengguna mengirim teks dan menerima hasil prediksi secara otomatis.

REST API bekerja berdasarkan endpoint. Endpoint adalah alamat URL yang menangani fungsi tertentu. Untuk analisis sentimen, endpoint utama biasanya adalah endpoint prediksi. Client mengirim teks ke endpoint tersebut, lalu server memprosesnya menggunakan model terlatih. Server kemudian mengembalikan hasil dalam bentuk terstruktur.

Endpoint prediksi umumnya menggunakan metode POST. Metode ini digunakan karena data teks dikirim dalam badan permintaan. Server menerima input, melakukan pra-pemrosesan, menjalankan model, dan mengirimkan respons. Proses ini berlangsung dalam waktu singkat agar sistem tetap responsif.



Contoh Postman API client

Format komunikasi biasanya menggunakan JSON. JSON mudah dibaca manusia dan mesin. Struktur ini ringan dan cocok untuk pertukaran data berbasis web. Konsistensi format sangat penting agar integrasi dengan sistem lain berjalan lancar.

Contoh format request sederhana dapat dituliskan sebagai berikut.

```
{
  "text": "Pelayanan sangat cepat dan ramah"
}
```

Server kemudian mengembalikan response dalam bentuk JSON seperti berikut.

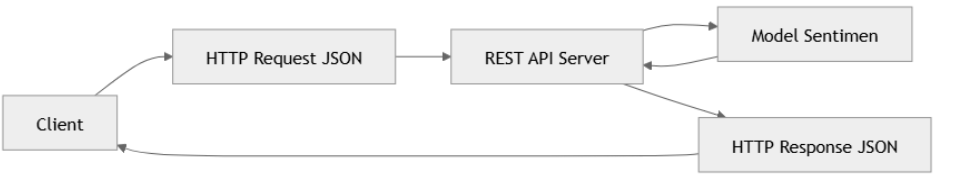
```
{
  "label": "positif",
  "probability": 0.92
}
```

Response dapat diperluas dengan informasi tambahan seperti waktu prediksi atau versi model. Struktur ini membantu proses monitoring dan audit sistem.

Keamanan menjadi aspek penting dalam implementasi REST API. Server perlu memvalidasi setiap input untuk mencegah serangan injeksi atau data berbahaya. Pembatasan ukuran teks juga dapat diterapkan untuk mencegah beban berlebihan. Autentikasi berbasis token sering digunakan untuk membatasi akses.

Validasi input memastikan bahwa teks tidak kosong dan sesuai format. Jika input tidak valid, server harus mengembalikan pesan kesalahan yang jelas. Penanganan kesalahan membantu menjaga stabilitas sistem. Log aktivitas juga perlu dicatat untuk kebutuhan pemantauan.

Struktur umum arsitektur REST API untuk analisis sentimen dapat digambarkan sebagai berikut.



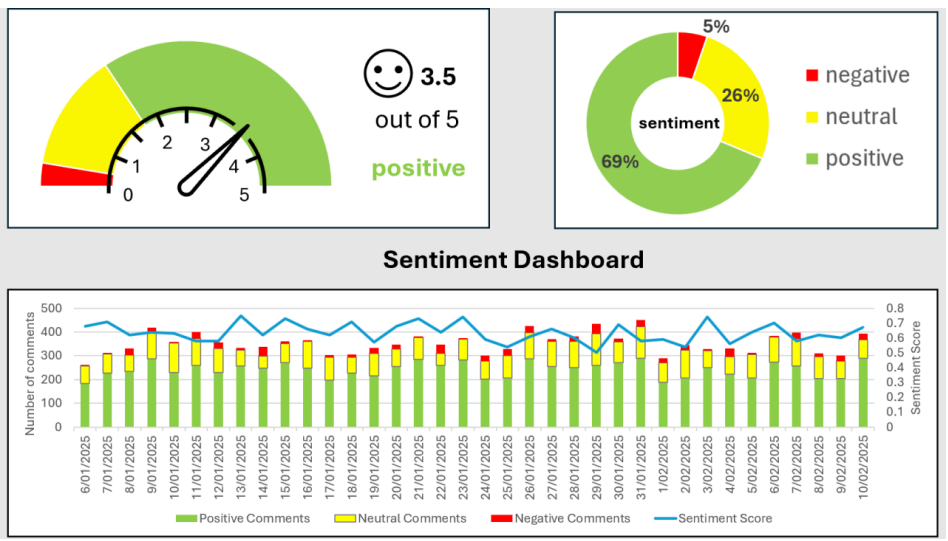
REST API memungkinkan integrasi fleksibel dengan berbagai aplikasi. Sistem e-commerce, media sosial, atau dashboard internal dapat memanfaatkan endpoint yang sama. Dengan desain yang baik, API menjadi fondasi komunikasi yang stabil dalam sistem analisis sentimen.

F. INTEGRASI DASHBOARD DAN VISUALISASI

Dashboard berfungsi menampilkan hasil analisis sentimen secara ringkas dan informatif. Visualisasi membantu pengguna memahami distribusi opini tanpa membaca seluruh data mentah. Dalam sistem produksi, dashboard menjadi antarmuka utama untuk monitoring. Desain yang jelas memudahkan interpretasi hasil.

Visualisasi hasil sentimen biasanya dimulai dengan ringkasan jumlah kelas. Sistem menghitung jumlah komentar positif, negatif, dan netral. Nilai tersebut kemudian ditampilkan dalam bentuk grafik batang atau pie chart. Visualisasi sederhana sudah cukup untuk memberikan gambaran umum.

Grafik distribusi sentimen membantu mengidentifikasi dominasi opini tertentu. Jika proporsi negatif meningkat, sistem dapat memberi peringatan dini. Distribusi ini juga berguna untuk evaluasi kampanye atau peluncuran produk baru. Perubahan distribusi mencerminkan respons publik.



Ilustrasi Tampilan Dashboard Analisis Sentimen

Monitoring tren waktu menambahkan dimensi temporal dalam analisis. Sistem mencatat tanggal atau waktu setiap input. Data ini dapat divisualisasikan dalam grafik garis. Grafik menunjukkan fluktuasi sentimen dari hari ke hari atau bulan ke bulan. Tren ini membantu memahami dinamika opini publik.

Contoh struktur monitoring tren waktu dapat digambarkan sebagai berikut.

Monitoring Tren Waktu

Integrasi dengan Streamlit memudahkan pembuatan dashboard interaktif. Pengguna dapat memasukkan teks dan melihat hasil prediksi langsung pada halaman yang sama. Streamlit juga menyediakan komponen grafik bawaan yang sederhana. Pendekatan ini cocok untuk prototipe atau aplikasi internal.

Dash menawarkan fleksibilitas lebih tinggi untuk visualisasi kompleks. Framework ini memungkinkan integrasi grafik interaktif berbasis web. Sistem dapat menampilkan filter waktu, kategori produk, atau wilayah tertentu. Dashboard menjadi alat analisis yang lebih dinamis.

Struktur umum dashboard analisis sentimen dapat dirangkum pada tabel berikut.

Komponen	Fungsi
Input Teks	Mengirim teks untuk dianalisis
Ringkasan Statistik	Menampilkan jumlah tiap kelas
Grafik Distribusi	Visualisasi positif, negatif, netral
Grafik Tren Waktu	Monitoring perubahan sentimen
Log Aktivitas	Menyimpan riwayat prediksi

Dashboard yang baik harus responsif dan mudah dipahami. Informasi utama perlu ditampilkan secara ringkas. Warna dan label harus konsisten agar tidak membingungkan pengguna. Integrasi visualisasi yang tepat membantu sistem analisis sentimen memberikan nilai praktis dalam pengambilan keputusan.

BAB 11. STUDI KASUS DAN APLIKASI AKADEMIK: DINAMIKA PSIKOLINGUISTIK DI ERA DIGITAL

Muhammad Oktoda Noorrohman S.Kom., M.Tr.T.

Transformasi digital dalam pendidikan tinggi di Indonesia bukan sekadar perpindahan platform dari ruang kelas ke layar komputer. Fenomena ini telah melahirkan sebuah ekosistem komunikasi yang sangat hidup sekaligus kompleks. Kini, opini, keluhan, hingga aspirasi mahasiswa tidak lagi tertahan di balik kuesioner formal, melainkan meluap bebas di "alun-alun digital" seperti X (Twitter), TikTok, dan YouTube. Di sinilah kejujuran emosional mahasiswa terekam secara mentah, jujur, dan tanpa filter (Handayani et al., 2020; Pramudita et al., 2024).

Menfess Sebagai Tempat Keluh Kesah

Namun, bagaimana kita memahami ribuan suara yang berseliweran tersebut? Di sinilah linguistik komputasi memainkan perannya melalui Analisis Sentimen (Sentiment Analysis). Metode ini bertindak seperti "penerjemah" yang membedah teks-teks tidak terstruktur menjadi wawasan strategis bagi para pengambil kebijakan di kampus maupun kementerian (Almutairi, 2025; Dinar et al., 2023). Tantangan terbesarnya adalah memahami karakteristik bahasa digital mahasiswa Indonesia yang sangat dinamis mulai dari ledakan bahasa gaul hingga sarkasme yang sangat halus (Bustamin et al., 2025; Iezzi, 2023). Untuk menangkap nuansa semantik yang utuh ini, kita telah berevolusi dari model leksikon tradisional menuju arsitektur Transformer yang canggih (Setiawan et al., 2025; Wilie et al., 2020).

A. LABIRIN LINGUISTIK: MEMAHAMI BAHASA DIGITAL GENERASI Z DAN ALPHA

Teks yang dihasilkan mahasiswa saat ini ibarat sebuah teka-teki linguistik. Di balik kalimat-kalimat cair mereka, tersimpan densitas informasi yang tinggi namun sering kali tersembunyi. Keberhasilan kita dalam menganalisis opini mereka sangat bergantung pada seberapa cerdas sistem kita membedah istilah-istilah baru seperti sat-set, yapping, hingga red flag yang telah menjadi standar komunikasi baru (Bustamin et al., 2025; Saputra et al., 2025). Istilah-istilah ini bukan sekadar tren, melainkan cerminan identitas sosial dan budaya yang berkembang pesat di era globalisasi (Mendrofa et al., 2024; Ningsih et al., 2024).

1. Dinamika Slang dan Glokalisasi Bahasa

Penggunaan bahasa gaul oleh Generasi Alpha dan Gen Z di ruang virtual menunjukkan fenomena "Glokalisasi"—adaptasi istilah internasional ke dalam konteks lokal (Pahruroji et al., 2025). Hal ini menciptakan variasi ejaan yang ekstrem, seperti penggunaan angka sebagai pengganti huruf (misalnya "s4ya" untuk "saya") atau pengulangan huruf yang berlebihan untuk menunjukkan intensitas emosi ("semangaaattt") (Faried et al., 2022). Karakteristik ini menambah beban kerja pada tahap pra-pemrosesan data karena mesin memerlukan kamus normalisasi yang selalu diperbarui (Khairul & Perdana, 2024).

2. Dinamika Alih Kode: Antara Prestise dan Kebiasaan

Di kota-kota besar seperti Jakarta dan Bandung, penggunaan alih kode (code-switching) antara bahasa Indonesia dan Inggris telah menjadi "napas"

komunikasi harian mahasiswa (Iezzi, 2023). Fenomena ini bukan sekadar penyimpangan aturan bahasa, melainkan strategi komunikasi fungsional untuk menunjukkan prestise, tingkat pendidikan, dan membangun kedekatan sosial (Skujins, 2024; Zalukhu et al., 2021).

Luca Iezzi (2023) menggambarkan fenomena ini sebagai "superimposition", di mana bahasa Inggris menyatu dalam struktur bahasa Indonesia tetap (Iezzi, 2023). Tantangan teknis muncul ketika sistem harus menentukan sentimen dari kalimat yang memiliki polaritas berbeda di setiap segmen bahasanya (Faried et al., 2022). Misalnya, kalimat "Tugas ini so difficult, tapi untungnya dosennya kind banget" mengandung muatan emosi yang kontras namun saling melengkapi.

3. Seni Mendeteksi Sarkasme Digital

Sarkasme adalah rintangan paling menantang dalam pemrosesan bahasa alami karena maknanya sering kali bertolak belakang dengan struktur kalimat aslinya (Suhartono et al., 2024). Mahasiswa sering menggunakan "senjata" ini untuk mengkritik kebijakan kampus tanpa harus terlihat menyerang secara langsung. Yunitasari et al. (2019) menyoroti bahwa keberadaan sarkasme bisa merusak akurasi data jika tidak ditangani secara khusus melalui proses pembalikan sentimen (sentiment reversal), di mana polaritas literal "positif" diubah menjadi "negatif" sesuai intensi asli pengguna (Yunitasari et al., 2019).

4. Analisis Multimodal: Melampaui Batas Tekstual

Dalam platform visual seperti TikTok dan Instagram, makna dalam opini mahasiswa tidak hanya dibangun melalui teks, tetapi juga melalui kombinasi simbol semiotik yang kompleks. Mahasiswa saat ini memandang platform tersebut sebagai ruang literasi multimodal yang menggabungkan elemen bahasa, auditori, dan visual secara utuh (Amin, 2023). Hasil pengamatan menunjukkan bahwa sentimen negatif atau tindakan perundungan (cyberbullying) sering kali diperkuat melalui pola penggunaan emoji tertentu, tanda baca yang dilebih-lebihkan, hingga pelapisan visual dalam komentar video (A. A. Sari et al., 2025). Oleh karena itu, sistem analisis sentimen masa depan harus mampu mengekstraksi fitur multimodal ini untuk menangkap intensitas emosi mahasiswa secara akurat (A. A. Sari et al., 2025).

B. DAPUR ANALISIS: METODOLOGI DAN EVOLUSI ALGORITMA

Untuk mengubah ribuan baris teks mentah dari media sosial menjadi data yang bermakna, kita memerlukan alur kerja yang sistematis. Bayangkan ini seperti mengolah bahan mentah menjadi hidangan yang siap saji bagi para pengambil keputusan (Yunitasari et al., 2019).

1. Arsitektur Pra-pemrosesan Tingkat Lanjut

Tahap ini krusial untuk membersihkan "kebisingan" (noise) agar model dapat belajar secara optimal. Langkah-langkah standar meliputi:

a. Case Folding: Menyeragamkan karakter menjadi huruf kecil untuk menghindari ambiguitas karakter (Yunitasari et al., 2019).

b. Tokenization: Memecah kalimat menjadi satuan kata atau token (Munthe et al., 2024).

c. Normalization: Mengubah kata slang atau singkatan menjadi bentuk baku menggunakan kamus khusus seperti Kamus Alay (Octavianus, 2021; Saputra et al., 2025).

d. Stopwords Removal: Menyingkirkan kata-kata umum yang tidak membawa muatan sentimen (Ningsih et al., 2024).

Metode pembobotan kata menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) tetap menjadi standar industri untuk memberikan nilai kepentingan pada sebuah kata dalam seluruh korpus data (Khaira et al., 2023).

2. Perbandingan Kinerja: NBC vs SVM

Peneliti di Indonesia sering membandingkan dua algoritma klasik ini. Naive Bayes (NBC) unggul dalam kecepatan dan kesederhanaan, sementara Support Vector Machine (SVM) lebih tangguh dalam menangani data teks yang tidak seimbang dan berdimensi tinggi (Khaira et al., 2023). Studi oleh Amandasari dan Damayanti (2025) menunjukkan keunggulan SVM dengan akurasi mencapai 79,32% dalam klasifikasi sentimen pelayanan publik, mengungguli NBC yang hanya mencapai 61,63% (Amandasari & Damayanti, 2025).

3. Revolusi Deep Learning dan IndoBERT

Evolusi terbaru dipimpin oleh model Transformer seperti IndoBERT. Dilatih menggunakan dataset Indo4B yang mencakup 4 miliar kata dari berita, media sosial, dan Wikipedia, IndoBERT mampu menangkap konteks kalimat secara dua arah (bidirectional) (Wilie et al., 2020). Dwi Setiawan (2025) membuktikan bahwa IndoBERT mampu mencapai akurasi maksimal sebesar 95% dalam menganalisis opini publik yang kompleks di YouTube, melampaui BERT tradisional dan Random Forest (Setiawan et al., 2025).

4. Transisi ke Model Emosi Majemuk: Teori Plutchik

Alih-alih hanya menggunakan klasifikasi biner positif-negatif, riset kontemporer mulai beralih ke model emosi majemuk Plutchik untuk mengidentifikasi label emosi yang lebih spesifik seperti anger, fear, joy, trust, hingga disapproval (Prastyo et al., 2025). Implementasi fitur emosi kelas jamak ini terbukti memberikan tingkat invarian yang signifikan dibandingkan dengan

sistem analisis sentimen tradisional, sehingga meningkatkan akurasi dan keberagaman analisis (Prastyo et al., 2025).

5. Arsitektur Dashboard dan Metodologi KDD

Pengembangan dashboard sistem pendukung keputusan di universitas memerlukan pendekatan metodologis Knowledge Discovery in Databases (KDD) untuk mengekstraksi dan menginterpretasikan wawasan signifikan dari basis data besar secara sistematis (Dinar et al., 2023). Untuk memastikan sistem yang dibangun optimal, penggunaan metodologi perancangan perangkat lunak iterative incremental sangat direkomendasikan guna memudahkan perancangan antarmuka dan modifikasi data secara berkelanjutan (Wijaya et al., 2023). Metodologi ini memungkinkan pengembangan fitur visualisasi grafik tren sentimen dilakukan secara bertahap dan terevaluasi dengan baik (Wijaya et al., 2023).

C. REALITAS DI LAPANGAN: INFRASTRUKTUR AKADEMIK DAN KEBIJAKAN

Aplikasi nyata analisis sentimen memberikan gambaran jujur mengenai apa yang sebenarnya dirasakan mahasiswa terhadap fasilitas pendidikan mereka.

1. Studi Kasus: Ketegangan SIMAK UNSRI

Masalah teknis pada Sistem Informasi Akademik (SIMAK) sering memicu ledakan emosi negatif. Zafira Thuraya (2025) mengungkapkan bahwa 64% sentimen mahasiswa terhadap SIMAK di Universitas Sriwijaya bersifat negatif (Thuraya, 2025). Keluhan utama berpusat pada kegagalan server saat masa pengisian KRS, informasi yang tidak lengkap, dan antarmuka yang kurang intuitif (Anggraini, 2021; D. A. Sari & Prabowo, 2020).

2. Evaluasi Program Kampus Merdeka dan KIP-Kuliah

Kebijakan nasional juga menjadi subjek pengawasan digital. Program Kampus Merdeka (MBKM) mendapatkan respons yang relatif positif dengan akurasi klasifikasi 97,92% menggunakan model NBC (Dinar et al., 2023). Sebaliknya, analisis terhadap program KIP-Kuliah menunjukkan dominasi sentimen negatif yang berkaitan dengan transparansi seleksi dan ketepatan sasaran bantuan (Pramudita et al., 2024).

D. KESEHATAN MENTAL: MENDETEKSI BURNOUT LEWAT JEJAK DIGITAL

Integrasi antara analisis sentimen dan psikologi mahasiswa kini menjadi kebutuhan mendesak di era tekanan digital yang tinggi.

1. Prevalensi Stres Akademik di Indonesia

Sebuah meta-analisis sistematis terhadap 63 penelitian menunjukkan realitas yang memprihatinkan: 72,3% mahasiswa Indonesia mengalami stres akademik kategori sedang hingga berat (Pratiwi et al., 2024; Zega & Zega, 2025). Beban akademik (r = 0,56) dan sistem evaluasi (r = 0,58) ditemukan sebagai kontributor terkuat terhadap tekanan ini (Zega & Zega, 2025).

2. Deteksi Gejala Depresi dan Burnout

Melalui teknik NLP, universitas kini dapat melakukan deteksi dini. Pola kata kunci seperti "stres", "lelah", "cemas", hingga frasa "burnout" menjadi indikator kuat (Dhinora & Mailoa, 2025; Rahmaddeni et al., 2025). Menariknya, mahasiswa mulai terbuka terhadap pemanfaatan Kecerdasan Buatan (AI) sebagai "cermin digital" untuk memantau kesehatan mental mereka, asalkan prinsip privasi dan transparansi algoritma terpenuhi (Wafiq et al., 2025).

E. ETIKA DATA DAN KEAMANAN INFORMASI MAHASISWA

Dalam melakukan pengumpulan data digital (scraping), isu etika menjadi sangat krusial. Mahasiswa sebagai subjek data memiliki hak atas privasi yang harus dilindungi.

1. Legalitas dan Tanggung Jawab Scraping

Web scraping secara hukum dianggap sah jika digunakan untuk kepentingan penelitian pribadi dan tidak melanggar undang-undang hak cipta (Sahria, 2022). Namun, peneliti harus berhati-hati dalam mempublikasikan ulang data opini atau ulasan pribadi tanpa anonimisasi yang tepat demi menghindari pelanggaran privasi (Jarmul, 2017).

2. Kesadaran Keamanan Informasi

Penelitian di Institut Teknologi Sepuluh Nopember (ITS) menunjukkan bahwa meskipun 95,3% mahasiswa sadar akan pentingnya keamanan informasi, masih diperlukan edukasi berkelanjutan mengenai penyebaran informasi pribadi di media sosial guna mencegah penyalahgunaan data oleh pihak tidak bertanggung jawab (Rakhmawati et al., 2024).

3. Implementasi UU PDP No. 27 Tahun 2022

Sebagai pengendali data, universitas wajib tunduk pada Undang -Undang Perlindungan Data Pribadi (UU PDP) Nomor 27 Tahun 2022. Regulasi ini mengatur kewajiban institusi dalam menjaga kerahasiaan data mahasiswa, di mana pelanggaran privasi dapat dikenai sanksi administratif hingga pidana (Abdullah et al., 2025). Pasal 67 UU PDP menegaskan bahwa pemanfaatan data pribadi milik orang lain secara melawan hukum dapat dikenai hukuman maksimal 5 tahun penjara dan/atau denda sebanyak Rp5 miliar (Abdullah et al., 2025). Tantangan besar saat ini adalah tingkat kepatuhan organisasi yang masih rendah; survei menunjukkan sekitar 60% perusahaan digital di Indonesia belum sepenuhnya menerapkan prinsip-prinsip perlindungan data (Nugroho & Ramadhani, 2023).

F. SINTESIS STRATEGIS: MENANGKAP "NYAWA" DARI ANGKA STATISTIK

Pada akhirnya, analisis sentimen bukan sekadar instrumen statistik yang dingin. Ia adalah jembatan komunikasi yang hangat antara institusi pendidikan dan civitas akademikanya (Almutairi, 2025). Dari data yang dikumpulkan, terlihat sebuah pola kritis: mahasiswa Indonesia kini lebih vokal di media sosial saat menghadapi hambatan teknis atau ketidakadilan sistemik (Dhinora & Mailoa, 2025; Thuraya, 2025).

Universitas harus segera mengadopsi dasbor analisis sentimen real-time sebagai Sistem Peringatan Dini (Early Warning System). Hal ini memungkinkan rektorat merespons cepat keluhan server saat pengisian KRS atau mendeteksi isu kesejahteraan mahasiswa sebelum berkembang menjadi krisis massal (Almutairi, 2025; D. A. Sari & Prabowo, 2020). Hasil riset mengenai stres akademik harus menjadi dasar bagi fakultas untuk meninjau kembali distribusi beban tugas guna mengurangi tingkat burnout mahasiswa (Zega & Zega, 2025).

Dengan memanfaatkan teknologi mutakhir seperti IndoBERT secara etis dan transparan, ekosistem pendidikan tinggi di Indonesia dapat menjadi lebih responsif dan empatik (Wafiq et al., 2025). Melalui integrasi data ini, institusi tidak hanya mendapatkan deretan angka statistik, tetapi juga menangkap

"nyawa" dari opini mahasiswa yang sebenarnya. Pendekatan berbasis data inilah yang akan memperkuat tata kelola universitas yang berpusat pada kepuasan pengguna, sekaligus meningkatkan kualitas output pendidikan nasional di kancah global (Almutairi, 2025).

DAFTAR PUSTAKA

Abdullah, C., Durand, N., & Moonti, R. M. (2025). Transformasi Digital dan Hak atas Privasi: Tinjauan Kritis Pelaksanaan UU Perlindungan Data Pribadi (PDP) Tahun 2022 di Era Big Data. Amandemen: Jurnal Ilmu Pertahanan, Politik Dan Hukum Indonesia, 2(3), 233–241. https://doi.org/10.62383/amandemen.v2i3.1073

Almutairi, M. (2025). Sentiment Analysis in Learning Management Systems: Understanding Student Feedback at Scale. ArXiv Preprint ArXiv:2506.05490. https://arxiv.org/abs/2506.05490

Amandasari, F., & Damayanti, D. (2025). Perbandingan Kinerja Support Vector Machine dan Naive Bayes dalam Klasifikasi Sentimen Twitter Terhadap Pelayanan BPJS. Jurnal Pendidikan Dan Teknologi Indonesia, 5(3), 645–653. https://doi.org/10.52436/1.jpti.680

Amin, F. M. (2023). Exploring Students' Reading Interest Through Tiktok Multimodal Literacy. Journal of Education and Teaching Learning (JETL), 5(2), 157–164. https://doi.org/10.51178/jetl.v5i2.1326

Anggraini, W. (2021). Analisis pada Sistem Informasi Akademik: Studi Usability Scale. Jurnal Saintek UNY.

Bustamin, A., Prayogi, A. A., Siswanto, D., Rafrin, M., & Nurdin, A. (2025). Text normalization for Indonesian slang words in sentiment analysis development. ICIC Express Letters, Part B: Applications, 16(2), 121–129. https://doi.org/10.24507/icicelb.16.02.121

Dhinora, M. Y., & Mailoa, E. (2025). Analisa Tweet Mahasiswa untuk Deteksi Gejala Depresi dengan Penerapan Natural Language Processing. Jurnal Indonesia: Manajemen Informatika Dan Komunikasi, 6(2), 1193–1211. https://doi.org/10.63447/jimik.v6i2.1405

Dinar, A. N., Irawan, A. S. Y., & Umaidah, Y. (2023). Analisis Sentimen Pada Pengguna Twitter Terhadap Program Kampus Merdeka Menggunakan Naïve Bayes. JATI (Jurnal Mahasiswa Teknik Informatika), 7(1), 755–762.

Faried, M. I., Tampubolon, L. P. D., & Atmodjo, D. (2022). Sentiment Analysis Research in Indonesian Language Reviewing From the Characteristics of

Comments. International Journal of Progressive Sciences and Technologies (IJPSAT), 32(2), 277–285.

Handayani, H., Luchia, N. T., Auliani, S. N., Azani, N. W., & Adha, R. (2020). Analysis of Twitter User Sentiments for the Aplikasi TikTok Application Using Naïve Bayes Clasifier Algorithm Method. SENTIMAS: Seminar Nasional Teknologi Informasi, Mekatronika Dan Ilmu Komputer.

Iezzi, L. (2023). Code switching online: the case of Indonesian and English on Instagram. Language. Text. Society, 10(2). https://ltsj.online/2023-10-2-iezzi

Jarmul, K. (2017). Python Web Scraping. Packt Publishing.

Khaira, U., Aryani, R., & Hardian, R. W. (2023). Perbandingan Algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam Analisis Sentimen Kebijakan Kemdikbudristek Tentang Kuota Internet Selama Pandemi Covid-19. PROCESSOR, 18(2), 183–195. https://doi.org/10.33998/processor.2023.18.2.897

Khairul, M. F. R., & Perdana, R. S. (2024). Arsitektur Sistem Percakapan Otomatis Berbahasa Indonesia dengan Normalisasi Bahasa Informal Menjadi Baku. Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK), 11(5), 1009–1016. https://doi.org/10.25126/jtiik.2024117984

Mendrofa, V. E. S., Ginting, L. E. B., Sitanggang, K., Nainggolan, M., Gaol, M. G. A. L., Hasibuan, S. E. F. B., & Purba, R. M. (2024). Bahasa Indonesia dan globalisasi: Kajian sosiolinguistik generasi alpha di tengah popularitas bahasa gaul (slang). Jurnal Ilmiah Kajian Multidisipliner, 8(9), 279–290.

Munthe, S. R., Suryadi, S., & Laksono, F. (2024). Analisis Klasifikasi Teks Pada Kata Slang di Media Sosial Menggunakan Pengolahan Bahasa Alami untuk Trending Topik. KLIK: Kajian Ilmiah Informatika Dan Komputer, 5(1), 230–241. https://doi.org/10.30865/klik.v5i1.2018

Ningsih, W., Alfianda, B., Rahmaddeni, R., & Wulandari, D. (2024). Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(2), 556–562. https://doi.org/10.57152/malcom.v4i2.1253

Nugroho, H., & Ramadhani, D. (2023). Perlindungan data pribadi dan tantangan implementasi UU PDP di era digital. Jurnal Kominfo & Hukum, 11(2), 75–90. https://doi.org/10.25077/jkh.v11i2.2023.75

Octavianus. (2021). Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based. JOINTER: Journal of Informatics Engineering, 2(02), 39–46.

Pahruroji, M., Mulyati, Y., & Cahyani, I. (2025). Glokalisasi bahasa: Dominasi dan adaptasi slang internasional pada kalangan generasi alpha di ruang virtual. Diglosia: Jurnal Kajian Bahasa, Sastra, Dan Pengajarannya, 8(4), 997–1010. https://doi.org/10.30872/diglosia.v8i4.1277

Pramudita, D., Akbar, Y., & Wahyudi, T. (2024). Analisis Sentimen Terhadap Program Kartu Indonesia Pintar Kuliah pada Media Sosial X Menggunakan Algoritma Naive Bayes. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(4), 1420–1430. https://doi.org/10.57152/malcom.v4i4.1565

Prastyo, D., Irawan, D., & Mursyidin, I. H. (2025). Analisis Sentimen Kelas Jamak Menggunakan Model Emosi Plutchik pada Data Media Sosial. BIT (Bina Insani ICT Journal), 12(1).

Pratiwi, Y., Handayani, L., & Sutanto, B. (2024). Prevalensi Stres Akademik Mahasiswa Indonesia: Studi Multi-Center. Jurnal Kesehatan Masyarakat, 19(1), 34–51.

Rahmaddeni, Fitria, V., Rahayu, K., Septhya, D., & Efrizoni, L. (2025). Deteksi Stres dan Depresi Unggahan Media Sosial dengan Machine Learning. JURNAL FASILKOM (Teknologi InFormASi Dan ILmu KOMputer), 15(1). https://doi.org/10.37859/jf.v15i1.9067

Rakhmawati, N. A., Rendiga, N. M., Aini, S. A., & Tertiabudi, V. A. (2024). Analisis Etika dalam Penggunaan Media Sosial Instagram Oleh Mahasiswa ITS Mengenai Pelanggaran Privasi. JIEET (Journal of Information Engineering and Educational Technology), 8(2), 90–95. https://doi.org/10.26740/jieet.v8n2.p90-95

Sahria. (2022). Implementasi Web Scraping Untuk Mengumpulkan Data Pada Media Sosial Instagram. Tugas Akhir. Universitas Islam Indonesia.

Saputra, A. N. A., Saputro, R. E., & Saputra, D. I. S. (2025). Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 9(2), 687–699. https://doi.org/10.33395/sinkron.v9i2.14613

Sari, A. A., Alya, H., Sihotang, R. O., Rafi, U., & Marisha, D. (2025). Multimodal Cyberbullying in TikTok Comments: A Descriptive Qualitative Analysis. 10(6), 1–10.

Sari, D. A., & Prabowo, H. (2020). Persepsi mahasiswa terhadap penggunaan sistem informasi akademik: Studi kasus pada Universitas Negeri di Jakarta. Jurnal Teknologi Informasi Dan Komunikasi, 8(1), 22–30.

Setiawan, V. D., Iswavigra, D. U., & Anggiratih, E. (2025). Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates. Scientific Journal of Informatics, 12(3), 397–406. https://doi.org/10.15294/sji.v12i3.26320

Skujins, C. E. A. (2024). Indonesian/English Code-Switching on Social Media: A Paper Exploring How to Mengekspresikan Diri Through the Switching of Bahasa. Master's Thesis. Flinders University.

Suhartono, D., Wongso, W., & Handoyo, A. T. (2024). IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection. IEEE Access, 12, 87323–87332. https://doi.org/10.1109/ACCESS.2024.3416955

Thuraya, Z. (2025). Analisis Sentimen Mahasiswa Terhadap Sistem Informasi Akademik Universitas Sriwijaya Menggunakan Metode Naive Bayes. Skripsi. Universitas Sriwijaya.

Wafiq, A., Syawal, A., Ikrar, & Ahmad, Z. A. (2025). Burnout Akademik di Era Digital dan Persepsi Mahasiswa terhadap Peluang Pemanfaatan Artificial Intelligence untuk Deteksi Dini. Jurnal Sains Dan Teknologi (JSIT), 5(3), 426–434. https://doi.org/10.47233/jsit.v5i3.3711

Wijaya, K. A., Romadhony, A., & Richasdy, D. (2023). Implementasi Model IndoBERT pada Dashboard Sentimen Media Sosial (Studi Kasus Universitas XYZ). EProceedings of Engineering, 10(4).

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 843–857. https://doi.org/10.18653/v1/2020.aacl-main.85

Yunitasari, Y., Musdholifah, A., & Sari, A. K. (2019). Sarcasm Detection For Sentiment Analysis in Indonesian Language Tweets. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(1), 53–62. https://doi.org/10.22146/ijccs.41136

Zalukhu, A. A. F., Laiya, R. E., & Laia, M. Y. (2021). Analysis of Indonesian-English Code Switching and Code Mixing on Facebook. Research on English Language Education, 3(2), 1–10. https://doi.org/10.57094/relation.v3i2.387

Zega, C. J. P., & Zega, F. P. (2025). Meta-Analisis Faktor-Faktor Yang Mempengaruhi Tingkat Stres Akademik Pada Mahasiswa. IDENTIK: Jurnal Ilmu Ekonomi, Pendidikan Dan Teknik, 2(2), 55–60. https://doi.org/10.70134/identik.v2i2.140

Buku ini mengajak pembaca memahami analisis sentimen secara menyeluruh, dari konsep dasar hingga implementasi sistem nyata. Pembahasan dimulai dari fondasi teoretis seperti linguistik komputasional dan teori emosi, lalu bergerak ke teknik representasi teks, machine learning, dan deep learning berbasis transformer. Setiap bab disusun sistematis agar pembaca memahami bukan hanya bagaimana model bekerja, tetapi juga mengapa pendekatan tertentu dipilih.

Materi dalam buku ini menggabungkan teori dan praktik. Pembaca akan mempelajari pra-pemrosesan teks, pembobotan fitur, evaluasi model, hingga deployment menggunakan REST API dan dashboard interaktif. Contoh tabel, diagram alur, dan ilustrasi arsitektur membantu menjembatani konsep dengan implementasi teknis. Topik lanjutan seperti fine-tuning model pre-trained dan penanganan data tidak seimbang juga dibahas secara aplikatif.

Buku ini cocok bagi mahasiswa, peneliti, dan praktisi yang ingin membangun sistem analisis sentimen secara mandiri. Pendekatan yang digunakan bersifat analitis, kontekstual, dan berbasis praktik nyata. Pembaca tidak hanya memahami algoritma, tetapi juga mampu merancang sistem yang akurat, stabil, dan siap digunakan dalam lingkungan produksi.

SAH PUBLISHER

PT Solusi Administrasi Hukum

Jl. Pedan Karangdowo, Klaten
publisher.sah.co.id
publisher@sah.co.id