BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian memanfaatkan teknik pemrosesan bahasa alami (NLP) dan pembelajaran mesin. Penelitian berfokus pada evaluasi dan klasifikasi dokumen sertifikasi halal. Fase awal merupakan gerbang masuk utama paling krusial dalam seluruh rantai pasok halal. Penelitian terinspirasi dari studi Kurniawati & Rochman (2023) tentang model optimasi distribusi produk halal menyoroti kompleksitas rantai pasok. Studi sebelumnya menekankan pentingnya sebuah sistem integrasi secara menyeluruh. Penelitian usulan menjawab kebutuhan dengan membangun fondasi kokoh dari berbagai bidang ilmu. Ekonomi halal, pemrosesan bahasa alami, dan pembelajaran mesin menjadi pilar utama dalam penelitian.

Mahardika et al. (2023) telah mendemonstrasikan nilai penerapan teknologi informasi dalam ekonomi halal dalam konteks lebih luas. Penelitian berhasil mengembangkan sistem rekomendasi kuliner halal berbasis lokasi untuk konsumen. Penelitian usulan sejalan dengan semangat pemanfaatan teknologi. Namun, penelitian memiliki orientasi berbeda dengan menargetkan pelaku usaha dan lembaga sertifikasi. Fokus penelitian adalah menciptakan efisiensi pada proses *back-office* sertifikasi. Proses *back-office* merupakan tulang punggung operasional bagi lembaga sertifikasi.

Kontribusi penting dari Kamari et al. (2025) dan Hasnan et al. (2024) memberikan inspirasi berharga bagi pendekatan penelitian. Karya menjadi dasar filosofis untuk menerapkan pendekatan sistematis dan berbasis bukti dalam proses klasifikasi dokumen. Pendekatan mensyaratkan berbagai faktor dan konteks harus dipertimbangkan secara komprehensif. Berbagai faktor kemudian diwakili oleh integrasi pengetahuan eksternal melalui mekanisme *Retrieval-Augmented Generation* (RAG). Kerangka kerja bertujuan meningkatkan akurasi dan keandalan sistem klasifikasi. Sistem pada akhirnya dapat mendukung proses pengambilan

keputusan lebih tepat.

Posisi dan kontribusi orisinal penelitian terletak pada pengusulan sebuah *Framework hybrid*. *Framework* secara spesifik mengintegrasikan teknik penanganan data tidak seimbang dengan model RAG untuk klasifikasi teks. Domain aplikasi penelitian berfokus pada dokumen sertifikasi halal. Kombinasi pendekatan merupakan *Novelty* signifikan dalam bidang. *Framework hybrid* dirancang untuk mengatasi dua tantangan utama secara simultan. Tantangan adalah masalah ketidakseimbangan data dan kebutuhan akan pemahaman kontekstual mendalam.

Tahap pertama dalam *Pipeline* adalah penerapan *Text Cleaning*. Proses berfungsi sebagai fondasi kritis dalam pra-pemrosesan data teks. He et al. (2024b) meneliti proses melibatkan eliminasi *noise*, normalisasi format, dan koreksi kesalahan secara sistematis. Pembersihan memastikan bahwa teks dari berbagai sumber dan format dokumen dapat diseragamkan dalam konteks sertifikasi halal. Proses standarisasi terjadi sebelum data diproses lebih lanjut. Pembersihan menyeluruh pada akhirnya meningkatkan konsistensi dan keandalan data secara signifikan.

Proses tokenisasi kemudian memecah teks mentah telah dibersihkan menjadi unit-unit diskrit. Unit-unit diskrit menjadi siap untuk diproses oleh mesin. Pendekatan berbasis aturan dan linguistik sangat diperlukan dalam tahap. Pendekatan khususnya menangani karakteristik unik dokumen hukum dan sertifikasi. Karakteristik unik itu mencakup singkatan khusus, istilah teknis, dan kata majemuk. Representasi teks sebagai himpunan token $\{w_1, w_2,..., w_n\}$ menjadi dasar fundamental. Seluruh analisis statistik dan pemodelan selanjutnya bergantung pada representasi dasar.

Langkah normalisasi *term*asuk *lowercasing* dan penghapusan simbol serta angka, menyamakan format data. Penyamarataan format memungkinkan model untuk memproses data dengan konsisten. He et al. (2024a) menyatakan bahwa normalisasi teks efektif dapat meningkatkan akurasi model secara signifikan. Peningkatan terjadi karena normalisasi mencegah kebingungan model.

Kebingungan sering kali muncul akibat perbedaan format atau gaya penulisan. Perbedaan format dan gaya penulisan itu sebenarnya tidak relevan dengan makna substantif sebuah dokumen.

Penghapusan *stopword* dilakukan untuk membersihkan teks dari kata-kata umum. Kata-kata umum tidak memberikan kontribusi signifikan terhadap makna keseluruhan dokumen. Pada konteks dokumen sertifikasi halal, kata seperti "dan", "di", atau "yang" biasanya dihilangkan. Proses penghapusan justru mempertahankan kata-kata teknis dan substantif mengandung informasi kritis. Informasi kritis sangat penting untuk menentukan kehalalan suatu produk atau proses. Hasil akhir dari proses adalah pengurangan dimensi data tanpa kehilangan informasi esensial dibutuhkan.

Stemming dan *lemmatization* lebih lanjut mengubah kata menjadi bentuk dasarnya. Proses sangat penting untuk menangani variasi morfologis dari istilah-istilah kunci dalam sertifikasi halal. Teknik memastikan bahwa kata seperti "disembelih", "menyembelih", dan "penyembelihan" memper*oleh* bentuk dasar sama. Model pengelompokan dan klasifikasi kemudian dapat mengenali berbagai variasi kata sebagai satu konsep terkait. Pengenalan terhadap konsep sama secara signifikan meningkatkan akurasi analisis. Peningkatan akurasi akhirnya mendukung kinerja keseluruhan sistem.

Data teks perlu diubah menjadi format numerik setelah teks dibersihkan dan dinormalisasi. Format numerik memungkinkan model pembelajaran mesin memproses data. Metode *Bag of Words* (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk membuat representasi vektor. Representasi vektor secara efektif menangkap esensi dari setiap dokumen. TF-IDF khususnya sangat efektif dalam memberikan bobot lebih tinggi pada istilah-istilah penting. Istilah-istilah spesifik untuk dokumen sertifikasi halal mendapatkan bobot lebih signifikan dibandingkan dengan kata-kata umum.

Namun, tantangan utama sering dihadapi dalam *dataset* dunia nyata adalah ketidakseimbangan kelas. Jumlah dokumen masuk kategori "lengkap" atau "disetujui" mungkin jauh lebih banyak daripada "perlu revisi" atau "ditolak".

Fenomena sangat umum terjadi dalam *dataset* sertifikasi halal. Di sinilah teknik *Over-sampling* seperti ADASYN diterapkan. Alabduallah et al. (2024) mengutip teknik secara dinamis menghasilkan sampel sintetis untuk kelas minoritas. ADASYN menyesuaikan generasi sampelnya berdasarkan distribusi data ada. Penerapan teknik meningkatkan sensitivitas model terhadap kelas jarang muncul.

Penelitian menyempurnakan hasil *sampling* dan membersihkan *noise* mungkin timbul dengan teknik *Under-sampling* seperti *Tomek links*. Teknik menghilangkan sampel-sampel ambigu dan berada di perbatasan antara kelas mayoritas dan minoritas. Prinsip sejalan dengan penelitian diusung Sakib et al. (2024). Kombinasi ADASYN dan *Tomek links* menciptakan *dataset* lebih seimbang. *Dataset* dihasilkan memiliki kualitas lebih tinggi. *Dataset* telah disempurnakan siap digunakan untuk proses pelatihan model.

Kebutuhan akan pemahaman kontekstual mendalam melampaui kemampuan model statistik biasa. Dokumen sertifikasi halal sering merujuk pada regulasi, standar, dan fatwa spesifik. Pemahaman terhadap dokumen memerlukan pengetahuan eksternal komprehensif. *Framework* penelitian mengintegrasikan model *Retrieval-Augmented Generation* (RAG) untuk mengatasi tantangan. Model RAG secara aktif mengambil informasi kontekstual relevan dari basis pengetahuan eksternal. Basis pengetahuan berisi regulasi halal, standar sertifikasi, dan panduan teknis terpercaya.

Pengetahuan diambil kemudian digabungkan dengan input dokumen asli. Proses penggabungan terjadi sebelum data akhirnya diproses oleh model *Generator*. Model *Generator* kemudian bertugas untuk membuat klasifikasi akhir berdasarkan informasi telah diperkaya. Mekanisme memastikan bahwa keputusan klasifikasi tidak hanya didasarkan pada pola statistik dalam data pelatihan. Keputusan didasarkan pada pemahaman mendalam terhadap aturan dan konteks substantif berlaku. Pendekatan sejalan dengan spirit penelitian Kamari et al. (2025) dalam bidang pengambilan keputusan multikriteria.

Integrasi RAG secara efektif menanamkan pengetahuan domain eksplisit ke dalam proses klasifikasi. Proses mengubah sistem dari sekadar pemodelan tekstual

menjadi sistem pendukung keputusan cerdas. Sistem dihasilkan sangat berbasis pada pengetahuan domain relevan. Hal secara langsung menjawab tantangan diidentifikasi dalam tinjauan pustaka. Sistem tidak hanya memprioritaskan akurasi klasifikasi tinggi. Sistem memastikan bahwa keputusan dihasilkan dapat dipertanggungjawabkan dan sesuai dengan prinsip-prinsip syariah.

Penelitian akan menerapkan berbagai model klasifikasi *robust* untuk menguji efektivitas *Framework hybrid*. Penelitian akan membandingkan kinerja model *Logistic Regression*, *Random Forest*, dan *XGBoost*. Perbandingan mengacu pada studi telah dilakukan oleh Hu et al. (2024). Eksperimen akan menunjukkan model mana paling sinergis dengan teknik *data balancing*. Kombinasi dengan teknik *augmentasi* pengetahuan menjadi faktor penilaian utama. Hasil perbandingan diharapkan dapat mengidentifikasi model terbaik untuk menyelesaikan masalah klasifikasi.

Penelitian tidak hanya menerapkan *state-of-the-art* dalam NLP dan *machine learning*. Penelitian melakukan integrasi inovatif dan kontekstual untuk domain sertifikasi halal. *Framework* diusulkan menjembatani *Gap* antara kemampuan pemodelan statistik murni dengan kebutuhan pemahaman domain mendalam. Pendekatan secara khusus dirancang untuk menangani masalah sangat spesifik dan kompleks. Integrasi menghasilkan sebuah sistem tidak hanya canggih secara teknis tetapi relevan secara praktis. Sistem pada akhirnya dapat memberikan dampak signifikan bagi para pelaku industri dan lembaga sertifikasi.

Kontribusi praktis dari penelitian adalah memberikan alat dapat membantu lembaga sertifikasi halal. Alat dirancang untuk memproses dokumen dengan lebih cepat, akurat, dan konsisten. Efisiensi dihasilkan dapat mempercepat seluruh rantai pasok halal. Penelitian mewujudkan ide diusung oleh Kurniawati & Rochman (2023). Percepatan dimulai dari gerbang paling awal dalam rantai pasok, yaitu proses sertifikasi. Dampak pada akhirnya mendorong pertumbuhan ekonomi halal lebih efisien dan terpercaya.

Penelitian memberikan kontribusi dengan mendemonstrasikan sebuah kombinasi teknik inovatif. Penelitian menggabungkan teknik NLP tradisional,

penanganan data tidak seimbang, dan arsitektur canggih seperti RAG. Kombinasi berhasil memecahkan masalah klasifikasi teks kompleks dan nyata dalam domain sertifikasi halal. Penelitian membuka jalan bagi penerapan *Framework* serupa di domain regulasi dan sertifikasi lainnya. Domain-domain biasanya membutuhkan pemahaman kontekstual sangat tinggi. *Framework* menjadi solusi untuk domain menghadapi tantangan ketidakseimbangan data serupa.

Tabel 2.1 *State of the Art*: Klasifikasi Dokumen Sertifikasi Halal dengan NLP dan ML

Peneliti (Tahun)	Fokus/Topik	Metode & Teknik yang Digunakan	Posisi & Kontribusi Orisinal Penelitian Usulan
Kurniawati & Rochman (2023)	Model Optimasi Distribusi Produk Halal	Studi konseptual tentang rantai pasok	Menjawab kebutuhan integrasi dengan membangun fondasi dari ekonomi halal, NLP, dan ML. Fokus pada fase awal (gerbang masuk) sertifikasi sebagai tulang punggung operasional.
Mahardika et al. (2022)	Penerapan TI dalam Ekonomi Halal (Konsumen)	Pengembangan Sistem Rekomendasi Kuliner Halal berbasis Lokasi	Orientasi berbeda: Menargetkan pelaku usaha & lembaga sertifikasi (*back-office*)
Kamari et al. (2025) & Hasnan et al.	Pendekatan Sistematis & Berbasis	Memberikan dasar filosofis	Mengadopsi pendekatan sistematis dengan mekanisme MCDM untuk mengintegrasikan

Peneliti (Tahun)	Fokus/Topik	Metode & Teknik yang Digunakan	Posisi & Kontribusi Orisinal Penelitian Usulan
(2024)	Bukti		pengetahuan eksternal.
He et al. (2024b)	Pra-Pemrosesan Teks	*Text Cleaning* (eliminasi *noise*, normalisasi format, koreksi kesalahan)	Menerapkan *Text Cleaning* sebagai tahap pertama dalam *Pipeline* untuk memastikan konsistensi dan keandalan data.
He et al. (2024a)	Normalisasi Teks	*Lowercasing*, penghapusan simbol & angka	Mengintegrasikan normalisasi teks (*low*ercasing, dll) ke dalam *Pipeline* pra-pemrosesan untuk dokumen sertifikasi.
Alabduallah et al. (2024)	Penanganan Data Tidak Seimbang	*Teknik Over-sampling* (ADASYN)	Menggunakan ADASYN untuk mengatasi masalah ketidakseimbangan kelas (contoh: "disetujui" vs "ditolak") yang umum dalam *dataset* sertifikasi.
Sakib et al. (2024)	Penanganan Data Tidak Seimbang	*Teknik Under-sampling* (*Tomek links*)	Menyempurnakan hasil *sampling* dengan mengkombinasikan bootstraping untuk menciptakan *dataset* yang lebih seimbang dan

Peneliti (Tahun)	Fokus/Topik	Metode & Teknik yang Digunakan	Posisi & Kontribusi Orisinal Penelitian Usulan
			berkualitas tinggi.
Hu et al. (2024)	Evaluasi Model Klasifikasi	Perbandingan model (*Logistic Regression, Random Forest, XGBoost*)	Menerapkan dan membandingkan model-model yang sama (LR, RF, XGBoost) untuk menguji efektivitas *framework hybrid* yang diusulkan.
Kamari et al. (2025)	Pengambilan Keputusan Multikriteria	Pendekatan sistematis	Sejalan dengan spirit penelitian mereka, tetapi diimplementasikan secara teknis melalui mekanisme RAG untuk memasukkan faktor kontekstual ke dalam klasifikasi.
Penelitian Usulan	Klasifikasi Dokumen Sertifikasi Halal	*Framework Hybrid*: NLP + Data *Balancing* + RAG	Kontribusi Orisinal: Mengusulkan *framework hybrid* yang mengintegrasikan teknik penanganan data tidak seimbang (ADASYN+*Tomek links*) dengan model RAG untuk klasifikasi teks dalam domain sertifikasi halal.

Peneliti (Tahun)	Fokus/Topik	Metode & Teknik yang Digunakan	Posisi & Kontribusi Orisinal Penelitian Usulan
			Novelty terletak pada kombinasi ini untuk mengatasi dua tantangan sekaligus: ketidakseimbangan data dan kebutuhan pemahaman kontekstual yang mendalam.

Penelitian ini mengusulkan sebuah *Pipeline* komprehensif untuk klasifikasi dokumen sertifikasi halal. *Pipeline* mengadopsi teknik *state-of-the-art* dalam pra-pemrosesan teks. Teknik disesuaikan dengan karakteristik unik dokumen sertifikasi. Proses vektorisasi menggunakan metode BoW dan TF-IDF untuk mengubah teks menjadi representasi numerik. Penanganan ketidakseimbangan data memanfaatkan kombinasi inovatif teknik ADASYN dan *Tomek links*. Augmentasi pengetahuan kontekstual mengintegrasikan model RAG untuk menarik pengetahuan eksternal. Proses klasifikasi menerapkan dan membandingkan berbagai model canggih seperti BERT. Evaluasi menyeluruh dilakukan untuk menemukan model paling sinergis dengan *framework hybrid*. Kontribusi praktis penelitian ini menghasilkan alat bantu bagi lembaga sertifikasi. Alat ini mampu memproses dokumen lebih cepat, akurat, dan konsisten.

2.2 Keaslian Penelitian

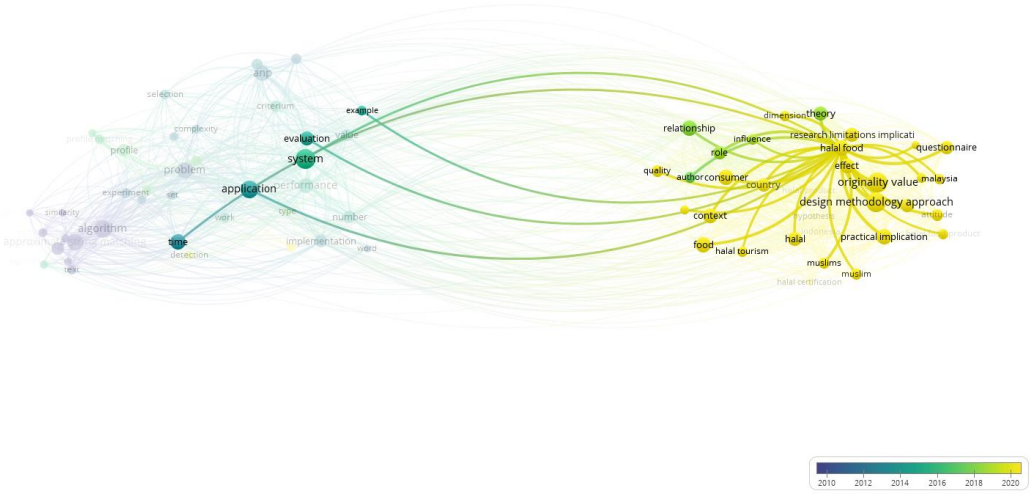
Hasil visualisasi peta bibliometrik menunjukkan pergeseran fokus penelitian signifikan dari tahun 2010 hingga 2020. Penelitian beralih dari aspek teknis menuju penerapan praktis dalam berbagai bidang. Studi berpusat pada pengembangan algoritma dan metode komputasi pada periode awal tahun 2010–2012. Para Penelitian banyak menggunakan kata kunci seperti *algorithm*,

approximation, dan *Similarity*. Kajian pada fase ini menekankan penguatan dasar-dasar teknis untuk membangun sistem. Penelitian berfokus pada penyelesaian masalah melalui pendekatan matematis dan komputasional. Fase awal ini menjadi fondasi penting bagi perkembangan selanjutnya. Para ilmuwan mengembangkan berbagai metode *detection* yang canggih selama periode ini. Dasar teknis yang kuat memungkinkan evolusi menuju aplikasi praktis di tahun-tahun berikutnya. Pergeseran ini mencerminkan pematangan bidang penelitian dari teori menuju praktik.

Penelitian mulai menunjukkan pergeseran fokus jelas pada periode pertengahan tahun 2014–2016. Para Penelitian mengalihkan perhatian pada implementasi dan evaluasi sistem nyata. Kata kunci seperti *system*, *Application*, dan *evaluation* mendominasi literatur penelitian selama periode ini. Kajian ilmiah mulai menguji coba berbagai algoritma yang dikembangkan sebelumnya. Penelitian lebih berfokus pada kinerja dan efektivitas sistem dalam kondisi riil. Para ilmuwan mengevaluasi implementasi teknologi dalam berbagai skenario aplikasi. Fase ini menandai transisi dari pengembangan teori menuju aplikasi praktis. Hasil pengembangan algoritma mulai diujicobakan dalam bentuk prototipe dan sistem kerja. Penelitian lebih menekankan pada kontribusi praktis teknologi bagi berbagai bidang. Periode pertengahan ini menjembatani kesenjangan antara inovasi teoritis dan penerapan dunia nyata.

Penelitian pada periode terbaru tahun 2018–2020 mengalami ekspansi ke ranah sosial-ekonomi lebih luas. Kajian kontemporer mulai menonjolkan kata kunci seperti halal *food*, halal *tourism*, dan halal *certification*. Sistem dan metodologi telah matang kemudian diterapkan dalam berbagai konteks industri halal. Penelitian mengangkat istilah seperti *consumer*, *role*, dan *relationship* secara signifikan. Fokus kajian bergeser pada perilaku konsumen dan dinamika pasar halal. Para Penelitian mengembangkan metodologi penelitian berbasis kuesioner dan studi lapangan. Kajian terkini lebih menekankan pada practical implication dan nilai orisinalitas. Tren ini menunjukkan perluasan aplikasi teknologi ke aspek sosial-budaya. Evolusi penelitian bergerak dari pendekatan teknis menuju multidisipliner

yang kontekstual. Periode terbaru ini menghasilkan kontribusi lebih aplikatif dan relevan bagi industri.



Gambar 2.1 Hasil Visualisasi *VOS Viewer* Penelitian Produk Makanan Halal

Penelitian memiliki keaslian dan *Novelty* jelas melalui beberapa aspek kunci. Penelitian mengusulkan integrasi belum lazim antara teknik penanganan data tidak seimbang dan model pemahaman bahasa mutakhir dari segi pendekatan teknis. *Framework hybrid* mengkombinasikan *Adaptive Synthetic Sampling* (ADASYN) dan *Tomek links* dengan model *Retrieval-Augmented Generation* (RAG) untuk klasifikasi teks. Kombinasi merupakan sebuah terobosan signifikan dalam bidang. Kebanyakan penelitian sebelumnya hanya berfokus pada salah satu aspek dengan menyeimbangkan data atau meningkatkan model. Penelitian-penelitian terdahulu belum menyatukan kedua pendekatan dalam sebuah *Pipeline* terintegrasi. *Pipeline* dirancang dalam penelitian berfungsi secara sinergis dan saling melengkapi. Sistem secara khusus dirancang untuk menyelesaikan kedua masalah utama secara simultan. Pendekatan holistik menjawab tantangan selama belum terselesaikan dalam klasifikasi dokumen kompleks. Keunikan integrasi menjadi nilai pembeda utama dari penelitian-penelitian sebelumnya.

Keaslian penelitian terletak pada konteks aplikasi dan domain spesifik ditargetkan. Komponen-komponen teknis seperti ADASYN, *Tomek links*, dan RAG memang telah diuji dalam berbagai domain lainnya. Domain-domain mencakup

deteksi penipuan atau analisis sentimen. Namun, penerapan teknik-teknik untuk klasifikasi dokumen sertifikasi halal masih sangat terbatas. Penelitian secara *Original* mengadaptasi dan memodifikasi teknik-teknik canggih. Adaptasi khusus dirancang untuk mengatasi karakteristik unik dari dokumen halal. Dokumen halal sendiri sarat dengan *term*inologi teknis dan konsep syariah kompleks. Dinamika regulasi terus diperbarui menjadi pertimbangan penting dalam adaptasi. Implementasinya bukan sekadar aplikasi langsung dari teknik sudah ada. Penelitian melakukan implementasi mempertimbangkan nuansa domain secara mendalam dan komprehensif.

Penelitian memberikan kontribusi metodologis *Original* melalui pendefinisian dan pembangunan *Knowledge Base* eksternal terstruktur. *Knowledge Base* secara khusus dirancang untuk mendukung mekanisme RAG secara optimal. Basis pengetahuan tidak hanya berisi regulasi statis umum digunakan. Tim Penelitian mengkurasi kontennya untuk mencakup daftar bahan diragukan kehalalannya. *Knowledge Base* memuat fatwa-fatwa terbaru dari otoritas kompeten. Berbagai studi kasus aktual dari lembaga sertifikasi terpercaya dimasukkan sebagai materi referensi. Konfigurasi spesifik untuk konteks halal merupakan sebuah inovasi kunci. Konfigurasi memungkinkan model RAG menghasilkan keputusan secara statistik akurat. Keputusan menjadi secara kontekstual relevan dengan prinsip-prinsip syariah. Aspek integrasi akurasi statistik dan relevansi syariah belum pernah dieksplorasi secara mendalam oleh penelitian-penelitian sebelumnya.

Aspek keaslian penelitian tampak dari pendekatan evaluasi komprehensif. Penelitian tidak hanya mengandalkan metrik akurasi tradisional. Metrik akurasi sering kali menyesatkan untuk evaluasi pada data tidak seimbang. Penelitian menggunakan kombinasi metrik *Precision*, *Recall*, dan *F1-score*. Setiap metrik diukur secara terpisah untuk setiap kelas ada. Selain itu, kerangka evaluasi mencakup perbandingan sangat. Perbandingan dilakukan antara model *hybrid* diusulkan dengan beberapa model *baseline*. Model *baseline* mencakup model tanpa penanganan data tidak seimbang dan model RAG *standalone*. Pembandingan diperlukan untuk membuktikan keunggulan setiap komponen

dalam *Framework* secara empiris.

Penelitian memiliki nilai orisinalitas dalam implikasi praktisnya langsung. Penelitian berbeda dari studi-studi sebelumnya di bidang serupa. Berbagai penelitian terdahulu seringkali hanya berakhir pada rekomendasi teoritis belaka. Penelitian justru dirancang untuk menghasilkan sebuah *prototype* model dapat diimplementasikan secara nyata. *Prototype* dapat diintegrasikan dengan portal sertifikasi halal daring telah ada. Solusi ditawarkan bukan hanya sebuah konsep akademis abstrak. Penelitian memberikan sebuah jawaban teknologi langsung dapat diadopsi oleh lembaga sertifikasi. Lembaga seperti LPPOM MUI dapat memanfaatkannya untuk mempercepat proses audit. Sistem mampu mengurangi beban kerja manual secara signifikan. Penelitian memperkuat ekosistem ekonomi halal Indonesia melalui automasi cerdas dan terpercaya.

2.3 Landasan Teori

2.3.1 Klasifikasi Teks dalam Pemrosesan Bahasa Alami

Klasifikasi teks merupakan bagian fundamental dalam bidang Pemrosesan Bahasa Alami (NLP). Tugas ini memetakan dokumen teks ke dalam kategori tertentu berdasarkan kontennya. Fernández et al. (2018) mendefinisikan klasifikasi teks sebagai proses formal pemetaan dokumen $d \in D$ ke dalam kelas $c \in C$. Dokumen sertifikasi halal menjadi domain aplikasi yang sangat relevan untuk tugas klasifikasi ini. Implementasi klasifikasi teks dalam domain ini menghadapi berbagai tantangan teknis yang kompleks.

Permasalahan ketidakseimbangan data sering muncul dalam *dataset* klasifikasi dokumen halal. Distribusi kelas yang tidak merata menyebabkan performa model klasifikasi menjadi bias. Model cenderung memiliki akurasi tinggi pada kelas mayoritas namun gagal mengenali kelas minoritas. Fenomena ini sangat riskan dalam konteks sertifikasi halal. Deteksi dini kandungan syubhat atau haram yang termasuk kelas minoritas justru paling krusial.

Teknik *sampling* muncul sebagai solusi untuk menangani masalah ketidakseimbangan data. Metode ini melakukan manipulasi distribusi data *training* untuk menyeimbangkan representasi kelas. *Adaptive Synthetic Sampling*

(ADASYN) merupakan pengembangan dari teknik Synthetic *sampling* konvensional. He, H., Bai, Y., Garcia, E., & Li (2008) mengemukakan bahwa ADASYN menghasilkan sampel sintetik secara adaptif berdasarkan tingkat kesulitan *learning*. Pendekatan ini lebih efektif dibandingkan teknik *oversampling* sederhana.

Proses generasi sampel sintetik dapat menimbulkan masalah baru dalam data. Tumpang tindih antara kelas mayoritas dan minoritas sering terjadi pasca penerapan Synthetic *sampling*. Tomek (1976) memperkenalkan *Tomek links* sebagai teknik pembersihan data. Metode ini mengidentifikasi dan menghapus instance yang membingungkan di perbatasan kelas. Implementasi *Tomek links* setelah Synthetic *sampling* mampu meningkatkan kualitas *dataset* secara signifikan.

Integrasi *Adaptive Synthetic Sampling* dengan *Tomek links* membentuk *framework hybrid* yang komprehensif. Kombinasi ini mengatasi ketidakseimbangan data sekaligus membersihkan *noise* di *decision boundary*. *Framework* menghasilkan *dataset training* yang lebih bersih dan seimbang. Model klasifikasi teks dapat belajar pola yang lebih jelas dari data hasil *processing*. Kinerja klasifikasi untuk semua kelas terutama kelas minoritas.

Retrieval-Augmented Generation (RAG) memberikan dimensi baru dalam pengolahan teks. Lewis et al. (2021) mendemonstrasikan kemampuan RAG dalam memperkaya representasi teks dengan informasi eksternal. Sistem RAG mengintegrasikan basis pengetahuan eksternal ke dalam proses *encoding* dokumen. RAG dapat mengakses database regulasi halal dan komposisi bahan pada konteks klasifikasi dokumen halal. Enrichment informasi ini sangat vital untuk akurasi klasifikasi.

Integrasi RAG dengan teknik *sampling* menciptakan *Pipeline processing* yang powerful. Sistem pertama-tama memperkaya representasi dokumen melalui mekanisme retrieval. Data yang telah diperkaya kemudian melalui proses balancing dengan *Adaptive Synthetic Sampling*. Tahap pembersihan data dengan *Tomek links* menjadi step akhir preparasi data. *Pipeline* ini memastikan model menerima input

yang kaya informasi dan terstruktur dengan baik. Kualitas data *training* yang optimal menjadi fondasi bagi pembangunan model klasifikasi yang *robust.*

Model klasifikasi teks modern seperti BERT mendapatkan manfaat besar dari *Pipeline* ini (Devlin et al., 2018). Representasi teks yang diperkaya RAG memudahkan model memahami konteks spesifik domain halal. Data *training* yang seimbang memungkinkan model belajar pola semua kelas secara merata. Clean *decision boundary* meminimalkan ambiguity dalam proses klasifikasi. Model dapat membuat keputusan yang lebih akurat dan terinformasi secara keseluruhan.

Evaluasi kinerja model menggunakan metrik-metrik standar klasifikasi teks. *Precision, recall,* dan *F1-score* menjadi indikator utama performa model. Perhatian khusus diberikan pada performa kelas minoritas yang kritis. Aplikasi pada *dataset* dokumen sertifikasi halal riil menguji efektivitas pendekatan ini. Hasil eksperimen menunjukkan peningkatan signifikan dibandingkan metode konvensional.

Kerangka teoritis ini membangun fondasi yang kuat untuk pengembangan sistem klasifikasi. Integrasi RAG dengan *advanced sampling Techniques* menawarkan solusi komprehensif. Pendekatan ini mengatasi keterbatasan data sekaligus meningkatkan kualitas representasi teks. Implementasi dalam domain sertifikasi halal membuktikan efektivitas *framework* yang diusulkan. Pengembangan lebih lanjut dapat menyesuaikan teknik-teknik ini untuk domain spesifik lainnya.

2.3.2 Teori Distribusi Data dan Analisis Ketidakseimbangan

2.3.2.1 Konsep Distribusi Data dalam *Machine learning*

Konsep distribusi data merupakan fondasi fundamental dalam *machine learning*. Fernández et al. (2018) mendefinisikan distribusi data sebagai pola penyebaran instance dalam ruang fitur yang menggambarkan karakteristik populasi data. Distribusi data mempengaruhi kemampuan model dalam mempelajari pola yang representatif dalam konteks klasifikasi teks. Distribusi yang ideal memungkinkan model mengenali pola semua kategori dengan baik. Sebaliknya, distribusi yang bermasalah dapat menyebabkan bias dalam prediksi model.

Distribusi merata atau *Balanced* merupakan kondisi ideal dalam *dataset*

machine learning (Fernández et al., 2018). Pada distribusi ini, semua kelas memiliki proporsi yang seimbang dan representatif. Model konvensional dapat bekerja optimal dengan data yang terdistribusi secara merata. Setiap kelas mendapatkan perhatian yang sama selama proses *training*. Akurasi model cenderung konsisten across semua kategori yang ada.

Distribusi merata memungkinkan model mempelajari karakteristik setiap kategori secara komprehensif dalam konteks klasifikasi dokumen halal. Model dapat mengenali pola dokumen "halal" dengan sama baiknya dengan pola dokumen "syubhat" atau "haram". Tidak ada kategori yang diabaikan selama proses pembelajaran. Generalisasi model terhadap data baru menjadi lebih baik. Performa model pun menjadi stabil dan dapat diandalkan.

Distribusi tidak merata atau *ImBalanced* merupakan masalah umum dalam *dataset* dunia nyata. Proporsi instance antar kelas menunjukkan perbedaan yang signifikan pada kondisi ini. Kelas mayoritas mendominasi *dataset* sementara kelas minoritas terwakili sangat sedikit. Masalah ini sangat prevalen dalam klasifikasi dokumen halal mengingat sebagian besar produk cenderung berstatus halal. Dokumen dengan status "haram" atau "syubhat" biasanya sangat jarang ditemui.

Distribusi *ImBalanced* menimbulkan berbagai konsekuensi serius bagi model klasifikasi. Model cenderung bias *toward* kelas mayoritas karena terekspos lebih banyak selama *training*. Akurasi untuk kelas minoritas menjadi sangat rendah meskipun *Overall accuracy* tinggi. Model mungkin mengklasifikasikan semua instance sebagai kelas mayoritas untuk memaksimalkan *accuracy*. Fenomena ini sangat berbahaya dalam konteks sertifikasi halal dimana deteksi status "haram" justru paling kritis.

Distribusi *ImBalanced* memperumit tugas model dalam domain klasifikasi teks dokumen halal. Fitur-fitur tekstual yang membedakan kategori minoritas menjadi sulit dipelajari. Model gagal membangun *decision boundary* yang tepat untuk memisahkan kategori langka. Representasi embedding untuk kategori minoritas menjadi tidak optimal. Model kesulitan menggeneralisasi pola untuk kategori-kategori yang jarang muncul.

Beberapa teknik khusus diperlukan untuk menangani masalah distribusi tidak merata. Re*sampling* methods seperti *oversampling* dan *undersampling* dapat menyeimbangkan distribusi data. Algorithm-level approaches menyesuaikan *loss function* atau threshold *decision*. *Ensemble* methods menggabungkan multiple *Classifier*s untuk meningkatkan performa. Data augmentation *Techniques* menghasilkan sampel sintetik untuk kategori minoritas.

Adaptive Synthetic Sampling (ADASYN) merupakan salah satu pendekatan *advanced* untuk menangani *ImBalanced* data. Teknik ini menghasilkan sampel sintetik secara adaptif berdasarkan tingkat kesulitan *learning*. Area *decision boundary* yang kompleks mendapatkan lebih banyak sampel sintetik. Pendekatan ini lebih efektif dibandingkan random *oversampling* konvensional. Model dapat mempelajari *decision boundary* yang lebih *robust*.

Pemahaman mendalam tentang distribusi data sangat krusial dalam pengembangan model klasifikasi teks. Analisis distribusi harus menjadi tahap awal dalam setiap proyek *machine learning*. Pemilihan teknik pre*processing* dan modeling harus mempertimbangkan karakteristik distribusi data. Evaluasi model pun harus menggunakan *metrics* yang *appropriate* untuk *ImBalanced* data. Pendekatan holistik ini memastikan pengembangan model yang *reliable* dan fair.

Integrasi pemahaman distribusi data dengan teknik-teknik *Handling ImBalanced* data membentuk landasan yang kuat. Pendekatan ini memungkinkan pengembangan model klasifikasi yang *robust* meskipun menghadapi data tidak seimbang. Pendekatan ini menjamin deteksi yang akurat untuk semua kategori dalam konteks klasifikasi dokumen halal. Model dapat mengenali pola baik yang common maupun yang rare dengan sama baiknya. Implementasi yang tepat akhirnya menghasilkan sistem sertifikasi yang lebih *reliable* dan *trustworthy*.

2.3.2.2 Teori *Long-tailed Distribution*

Teori *long-tailed distribution* menjelaskan fenomena umum dalam data real-world. Data sebagian kecil kelas mendominasi sebagian besar instance. Y. Yang et al. (2021) mengemukakan bahwa distribusi ini ditandai dengan ketidakseimbangan ekstrem antara *head Classes* dan *tail Classes*. Dalam konteks

klasifikasi dokumen halal, distribusi ini muncul karena sebagian besar produk tergolong halal sementara kategori haram dan syubhat sangat jarang. *Head Classes* merepresentasikan kategori umum seperti "halal" dengan instance yang melimpah. *Tail Classes* mencakup kategori langka seperti "haram" atau "syubhat" dengan instance terbatas.

Karakteristik fundamental *long-tailed distribution* terletak pada struktur hierarkis kelasnya. *Head Classes* terdiri dari sejumlah kecil kelas dengan banyak instance yang mendominasi *dataset*. *Tail Classes* mencakup banyak kelas dengan sedikit instance yang tersebar secara *sparse*. ImBalance ratio dapat mencapai 1000:1 pada *dataset* ekstrem. Kondisi ini menciptakan tantangan *learning* yang signifikan. ImBalance ratio yang tinggi menyebabkan model kesulitan mempelajari representasi *tail Classes* dalam klasifikasi dokumen halal. Performa model untuk kategori kritis justru menjadi paling buruk.

Distribusi *long-tailed* dapat dimodelkan secara matematis menggunakan power law *distribution* (Y. Yang et al., 2021). Formulasi matematisnya dinyatakan sebagai persamaan 2.1.

$$P(y = c) = \frac{1}{c^{\alpha}} \quad \text{untuk } c = 1,2,\ldots,C \tag{2.1}$$

α merupakan parameter yang mengontrol derajat ketidakseimbangan. Nilai α yang lebih besar menunjukkan ketidakseimbangan yang lebih ekstrem dalam distribusi kelas. Parameter C merepresentasikan total jumlah kelas dalam klasifikasi. Formula ini menggambarkan probabilitas suatu instance termasuk dalam kelas c menurun secara eksponensial. Nilai α yang tinggi mencerminkan dominasi kuat kategori "halal" atas kategori lainnya pada *dataset* dokumen halal.

Head Classes dalam distribusi *long-tailed* memainkan peran dominan dalam proses *training* model. Kelas-kelas ini memiliki representasi data yang cukup untuk mempelajari pola yang komprehensif. Model dapat dengan mudah mengenali karakteristik *head Classes* karena *exposure* yang tinggi selama *training*. Namun, dominasi *head Classes* menyebabkan model mengembangkan bias yang kuat terhadap kelas mayoritas. Kategori "halal" sebagai *head Class* mendapatkan

perhatian berlebihan dari model dalam konteks sertifikasi halal.

Tail Classes menghadapi tantangan fundamental dalam *long-tailed distribution*. Kelas-kelas ini memiliki data *training* yang tidak memadai untuk mempelajari representasi yang bermakna. Model kesulitan membangun *decision boundary* yang tepat untuk memisahkan *tail Classes* (Kubat & Matwin, 1997). Generalisasi terhadap instance baru dari *tail Classes* menjadi sangat lemah. Padahal, dalam klasifikasi dokumen halal, *tail Classes* justru mengandung kategori paling kritis yang memerlukan deteksi akurat.

ImBalance ratio yang ekstrem memperparah masalah *long-tailed distribution* (Y. Yang et al., 2021). Rasio antara instance terbanyak dan tersedikit dapat mencapai persamaan 2.2.

$$\text{IR} = \frac{N_{max}}{N_{min}} \tag{2.2}$$

N_{max} merupakan jumlah instance *head Class* dan N_{min} adalah jumlah instance tail Class. Rasio tinggi menyebabkan *loss function* didominasi oleh *head Classes* selama *Optimization*. *Gradient* update terutama dipengaruhi oleh kelas mayoritas, mengabaikan pembelajaran *tail Classes*. ImBalance ratio pada *dataset* halal dapat sangat variatif tergantung kompleksitas produk dalam praktik.

Pendekatan konvensional menghadapi keterbatasan serius dalam menangani *long-tailed distribution*. Standard cross-entropy *loss* cenderung bias *toward head Classes* karena dominasi numerik. *Accuracy metrics* menjadi misleading karena kinerja *tail Classes* yang buruk tertutupi performa *head Classes*. Data splitting konvensional dapat menghasilkan *fold Validation* yang tidak merepresentasikan *tail Classes*. Evaluasi model pun memerlukan *metrics* khusus seperti *Balanced accuracy* atau *F1-score* macro.

Berbagai teknik *advanced* telah dikembangkan untuk mengatasi tantangan *long-tailed distribution* (Silverman, 1986). *Re-Weighting* methods menyesuaikan bobot *loss function* berdasarkan frekuensi kelas. Margin-based methods menerapkan *Large*r margins untuk *tail Classes* selama *training*. Two-stage *training* memisahkan fase *Feature learning* dan *Classifier adjustment*. *Transfer learning*

memanfaatkan pengetahuan dari *head Classes* untuk memperkuat *learning tail Classes*.

Dalam konteks klasifikasi teks dokumen halal, pendekatan *hybrid* diperlukan untuk menangani *long-tailed distribution*. *Adaptive Synthetic Sampling* (ADASYN) dapat menghasilkan sampel sintetik untuk *tail Classes* (He, H., Bai, Y., Garcia, E., & Li, 2008). *Ensemble learning* menggabungkan multiple specialized *Classifier*s untuk *different* parts of the *distribution*. *Cost-sensitive learning* menetapkan higher *misclassification* cost untuk *tail Classes*. *Knowledge* distillation memanfaatkan model yang dilatih pada data seimbang untuk memandu pembelajaran.

Pemahaman komprehensif tentang *long-tailed distribution* essential untuk pengembangan model klasifikasi yang *robust*. Analisis distribusi harus menjadi foundational step dalam *Pipeline* development. Pemilihan *Techniques* harus mempertimbangkan karakteristik spesifik *long-tailed distribution* pada domain halal. Evaluasi model perlu menggunakan *comprehensive metrics* yang sensitif terhadap performa *tail Classes*. Implementasi yang tepat menjamin sistem klasifikasi yang *reliable* untuk semua kategori, baik head maupun *tail Classes*.

2.3.2.3 Ukuran dan Metrik Ketidakseimbangan Data

Pengukuran ketidakseimbangan data merupakan langkah kritis dalam analisis *dataset* untuk klasifikasi teks. Cao et al. (2019) mengidentifikasi berbagai metrik kuantitatif yang dapat mengukur tingkat ketidakseimbangan secara objektif. Metrik-metrik ini memberikan dasar numerik untuk mengevaluasi parahnya masalah *ImBalanced* data dalam *dataset* dokumen halal. Pemahaman yang tepat terhadap ukuran ketidakseimbangan memungkinkan peneliti memilih teknik penanganan yang sesuai. Implementasi metrik yang komprehensif membantu dalam memonitor efektivitas metode balancing yang diterapkan.

ImBalance Ratio (IR) merupakan metrik fundamental yang paling umum digunakan dalam mengukur ketidakseimbangan data (Cao et al., 2019). Metrik ini didefinisikan secara matematis sebagai persamaan 2.3.

$$IR = \frac{\max_i(n_i)}{\min_j(n_j)} \quad (2.3)$$

n_i merupakan jumlah sampel kelas i. IR mengkuantifikasi rasio antara kelas mayoritas dan kelas minoritas dalam *dataset*. Nilai IR = 1 menunjukkan *dataset* yang perfectly *Balanced*, sementara nilai IR > 1 mengindikasikan ketidakseimbangan. Dalam konteks dokumen halal, nilai IR yang tinggi mencerminkan dominasi ekstrem kategori "halal" terhadap kategori lainnya.

Degree of ImBalance (DI) memberikan perspektif alternatif dalam mengukur ketidakseimbangan data (Japkowicz & Stephen, 2002). Formula matematis untuk DI adalah persamaan 2.4.

$$DI = 1 - \frac{\min_i(n_i)}{\max_j(n_j)} \quad (2.4)$$

DI mengukur proporsi ketidakseimbangan dalam rentang 0 hingga 1. DI = 0 menunjukkan keseimbangan sempurna. Nilai DI yang mendekati 1 mengindikasikan ketidakseimbangan yang sangat parah. Metrik ini sangat sensitif terhadap keberadaan kelas dengan instance yang sangat sedikit. Pada *dataset* sertifikasi halal, DI membantu mengidentifikasi sejauh mana kelas minoritas terabaikan.

Gini Coefficient untuk ketidakseimbangan kelas memberikan pengukuran yang lebih komprehensif mengenai distribusi data. Formula matematis dari Gini Coefficient dinyatakan sebagai Persamaan 2.5 (Fernández et al., 2018). C merupakan jumlah total kelas. Metrik ini secara khusus mengukur ketimpangan distribusi instance di semua kelas dalam suatu *dataset*. Nilai G = 0 menunjukkan distribusi yang perfectly equal, sementara nilai G = 1 mengindikasikan ketidakseimbangan sempurna. Metrik ini sangat efektif ketika diaplikasikan pada *dataset* dengan multiple Classes yang mengalami ketidakseimbangan kompleks karena kemampuannya untuk menangkap variasi di antara semua pasangan kelas.

$$G = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C}|n_i - n_j|}{2C\sum_{i=1}^{C} n_i} \quad (2.5)$$

Setiap metrik ketidakseimbangan memiliki keunggulan dan keterbatasan tersendiri dalam aplikasi praktis. ImBalance Ratio mudah diinterpretasikan namun

sensitif terhadap *outlier* dan kelas dengan satu instance. Degree of ImBalance memberikan normalisasi yang baik namun kurang informatif untuk multi-Class scenarios (Fernández et al., 2018). Gini Coefficient komprehensif namun Computationally lebih intensif untuk *dataset* besar. Pemilihan metrik yang tepat bergantung pada karakteristik spesifik *dataset* dan tujuan analisis.

Penerapan metrik ketidakseimbangan mengungkap tantangan mendalam yang dihadapi dalam konteks klasifikasi dokumen halal. Perhitungan IR biasanya menunjukkan nilai yang sangat tinggi mengingat jarangnya produk haram dibandingkan halal. DI mendekati 1 mencerminkan kesenjangan yang besar antara kelas mayoritas dan minoritas. Gini Coefficient yang tinggi mengkonfirmasi ketimpangan distribusi across semua kategori sertifikasi. Analisis metrik ini membuktikan kebutuhan urgent untuk teknik balancing yang efektif.

Interpretasi hasil pengukuran ketidakseimbangan memerlukan pertimbangan domain *Knowledge* yang mendalam. Nilai IR = 100 tidak selalu bermakna sama untuk berbagai konteks aplikasi. Ketidakseimbangan tinggi memang Expected mengingat nature industri makanan yang didominasi produk halal dalam sertifikasi halal. Namun, deteksi yang akurat untuk kelas minoritas tetap critical meskipun frekuensinya rendah. Ambang batas untuk menentukan "ketidakseimbangan parah" perlu disesuaikan dengan konteks spesifik.

Metrik ketidakseimbangan berperan penting dalam evaluasi efektivitas teknik balancing. Penurunan nilai IR, DI, dan Gini Coefficient setelah penerapan *sampling Techniques* mengindikasikan keberhasilan balancing. Namun, *improvement* dalam metrik ketidakseimbangan tidak selalu berkorelasi langsung dengan peningkatan performa klasifikasi. Over-*sampling* yang agresif dapat menurunkan ketidakseimbangan metrik namun menyebabkan *overfitting*. Metrik ketidakseimbangan harus digunakan bersamaan dengan metrik evaluasi model lainnya.

Integrasi metrik ketidakseimbangan dalam *Pipeline machine learning* memungkinkan pengembangan sistem yang lebih *robust*. Automated monitoring terhadap metrik ketidakseimbangan dapat mendeteksi data drift dalam sistem

production. Dynamic *adjustment* of *sampling* strategies berdasarkan nilai metrik ketidakseimbangan meningkatkan adaptabilitas sistem. Alert mechanisms dapat diaktifkan ketika metrik ketidakseimbangan melebihi threshold tertentu. Pendekatan proactive ini menjamin maintainability sistem klasifikasi dalam jangka panjang.

2.3.2.4 Dampak Distribusi Tidak Seimbang pada Model

Distribusi tidak seimbang dalam *dataset* klasifikasi teks menimbulkan dampak sistemik pada proses pembelajaran model *machine learning*. Krawczyk (2023) menganalisis secara komprehensif berbagai dampak negatif yang muncul akibat ketidakseimbangan data. Dampak-dampak ini tidak hanya mempengaruhi performa model secara superficial tetapi merusak fundamental *learning* process itu sendiri. Dampak ketidakseimbangan dapat berakibat fatal mengingat pentingnya deteksi akurat untuk kategori minoritas dalam konteks klasifikasi dokumen halal. Pemahaman mendalam tentang mekanisme dampak ini essential untuk pengembangan teknik mitigasi yang efektif.

Bias *Gradient updates* merupakan dampak fundamental yang langsung mempengaruhi proses optimasi model. Update rule dinyatakan sebagai persamaan 2.6 dalam stochastic *Gradient* descent (Fernández et al., 2018).

$$\nabla \mathcal{L} \approx \frac{1}{N} \sum_{i=1}^{N} \nabla \mathcal{L}(x_i, y_i) \tag{2.6}$$

N merupakan total jumlah sampel. Pada data tidak seimbang, *Gradient* didominasi oleh kontribusi kelas mayoritas karena N*majority* >> Nminority. Akibatnya, parameter model diupdate terutama untuk meminimalkan *loss* pada kelas mayoritas. Konvergensi model pun menjadi bias *toward* kelas mayoritas. Model mengabaikan pembelajaran pola kelas minoritas yang justru kritis dalam sertifikasi halal.

Decision boundary shift merupakan fenomena dimana model secara sistematis menggeser batas keputusan menjauhi kelas minoritas. Model mengoptimalkan *decision boundary* untuk memaksimalkan akurasi *Overall* dengan mengorbankan kelas minoritas. Secara matematis dimodelkan sebagai pergeseran

hyperplane *decision* pada persamaan 2.7 (Fernández et al., 2018).

$$w^T x + b = 0 \rightarrow w^T x + b' = 0, \text{ dimana } b' > b \qquad (2.7)$$

Pergeseran ini menyebabkan region *decision* untuk kelas minoritas menyusut secara signifikan. Dalam klasifikasi dokumen halal, pergeseran *boundary* menyebabkan under-detection untuk kategori "syubhat" dan "haram" yang berisiko tinggi.

Feature learning bias terjadi ketika model mempelajari representasi fitur yang tidak representative untuk semua kelas. Model lebih fokus mempelajari Feature yang *discriminative* untuk kelas mayoritas pada data tidak seimbang. Embedding space yang dihasilkan menjadi terdistorsi dimana instance kelas mayoritas memiliki representasi yang kaya sementara kelas minoritas terkompresi (Fernández et al., 2018). Secara formal dinyatakan sebagai persamaan 2.8.

$\phi(x) = f_\theta(x)$ dimana θ dioptimasi untuk max P(y*majority* | ϕ(x)) (2.8)

Akibatnya, model kesulitan membedakan instance kelas minoritas dalam embedding space yang bias tersebut. Dampak ketidakseimbangan pada *evaluation metrics* menciptakan illusion of good Performance.

Overfitting pada kelas mayoritas merupakan konsekuensi lain dari distribusi tidak seimbang (Fernández et al., 2018). Model menjadi overspecialized dalam mempelajari pola kelas mayoritas yang berlimpah data. Secara matematis, ini tercermin dalam variance of predictions dengan kondisi Var[$\hat{y}$*majority*] < Var[ŷminority].

Model mengembangkan kapasitas berlebihan untuk mengenali variasi dalam kelas mayoritas sementara mengabaikan kelas minoritas (Krawczyk, 2018). *Regularization Techniques* konvensional sering gagal mengatasi masalah ini karena tidak explicitly address ImBalance dalam data.

Generalization capability model menjadi terbatas akibat distribusi tidak seimbang (J. M. Johnson & Khoshgoftaar, 2019). Model yang dilatih pada data ImBalance menunjukkan performa buruk ketika dihadapkan pada distribusi data real-world yang mungkin berbeda. Ketidakmampuan generalisasi ini particularly critical dalam klasifikasi dokumen halal pada pola dokumen baru terus bermunculan. Model gagal mengembangkan *robust* representation yang diperlukan

untuk menangani variasi dalam kategori minoritas. *Transfer learning* pun menjadi tidak efektif karena foundational representation yang bias.

Model terpengaruh secara signifikan oleh ketidakseimbangan data (Fernández et al., 2018). Model cenderung *overconfident* dalam memprediksi kelas mayoritas dan *underconfident* untuk kelas minoritas. Secara probabilistik, ini dapat dinyatakan sebagai P(ŷ = y*majority* | x) >> True *Probability* dan P(ŷ = yminority | x) << True *Probability*.

Ketidakakuratan calibration ini berbahaya dalam *decision making* untuk sertifikasi halal dimana confidence score digunakan sebagai basis approval. Sistem mungkin secara keliru memberikan confidence tinggi untuk prediksi yang salah pada kelas mayoritas.

Dampak kumulatif dari berbagai efek di atas menghasilkan model yang *fundamentally flawed* untuk aplikasi praktis. Model tidak dapat dipercaya untuk deteksi kategori kritis meskipun performa *Overall metrics* tampak baik. Dalam industri halal certification, konsekuensinya dapat berupa produk haram yang lolos sertifikasi atau produk halal yang ditolak tanpa alasan jelas. Biaya *False* Negative (gagal mendeteksi haram) jauh lebih tinggi daripada *False Positive* dalam konteks ini. Mitigasi ketidakseimbangan bukan hanya masalah teknikal tetapi *ethical* imperative.

Pemahaman komprehensif tentang dampak distribusi tidak seimbang membuka jalan untuk pengembangan solusi yang *targeted* dan efektif. Teknik seperti *Cost-sensitive learning* langsung address bias dalam *Gradient updates*. *Sampling* methods mengatasi ImBalance pada level data *distribution* sebelum *training*. *Ensemble* methods combining multiple specialized models dapat mengatasi *Feature learning* bias. *Evaluation metrics* yang *appropriate* mencegah misinterpretasi performa model. Pendekatan holistic ini essential untuk membangun sistem klasifikasi dokumen halal yang *reliable* dan fair.

2.3.3 *Transformer Architecture* dan *Pre-trained Language Models*

2.3.3.1 *Transformer Architecture*

Arsitektur Transformer merevolusi bidang pemrosesan bahasa alami dengan mengatasi keterbatasan model sequential sebelumnya. Vaswani et al. (2017) memperkenalkan arsitektur berbasis perhatian (*attention*) yang sepenuhnya mengeliminasi ketergantungan pada recurrent dan *convolutional* Layers. Transformer memanfaatkan mekanisme *self-attention* untuk memproses seluruh *sequence* secara paralel, sehingga model mempercepat waktu pelatihan. Arsitektur ini memungkinkan pemodelan ketergantungan jangka panjang (long-*Range dependencies*) dalam teks yang sangat relevan untuk dokumen halal yang kompleks. Kemampuannya menangkap konteks global menjadikan Transformer pilihan ideal untuk tugas klasifikasi teks dokumen teknis.

Mekanisme *self-attention* merupakan jantung dari arsitektur Transformer yang memungkinkan setiap token dalam *sequence* untuk berinteraksi dengan semua token lainnya. Secara matematis, mekanisme *attention* dihitung menggunakan fungsi scaled dot-product *attention* pada persamaan 2.9 (Vaswani et al., 2017).

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (2.9)$$

Q (Query), K (Key), dan V (Value) merupakan transformasi linear dari input embeddings. Pembagian dengan $\sqrt{d_k}$ berfungsi untuk menstabilkan gradien selama pelatihan, terutama untuk dimensi yang besar. Mekanisme ini memungkinkan model mengaitkan kata seperti "enzim" dengan "sumber" dan "hewan" meskipun posisinya berjauhan dalam konteks dokumen halal. Kemampuan menangkap relasi semantik kompleks ini sangat krusial untuk memahami spesifikasi produk halal.

Multi-head attention memperluas kapasitas model dengan menjalankan beberapa mekanisme *attention* secara paralel. Formulasi *Multi-head attention* didefinisikan sebagai persamaan 2.10 (Vaswani et al., 2017).

$$\text{Multi}Head(Q, K, V) = \text{Concat}(head_1, \dots, head_h)\, W^O \quad (2.10)$$

Setiap $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ (head *attention*) dapat mempelajari representasi yang berbeda dari dependensi dalam data. Misalnya, satu head dapat

fokus pada relasi sintaksis sementara head lain menangkap relasi semantik. *Multi-head attention* memungkinkan penangkapan berbagai aspek seperti komposisi bahan, proses produksi, dan sumber bahan secara simultan untuk klasifikasi dokumen halal.

Arsitektur Transformer mengintegrasikan komponen penting seperti positional *encoding* dan *feed-forward networks* (Vaswani et al., 2017). Positional *encoding* ditambahkan ke input embeddings untuk memberikan informasi mengenai posisi token dalam *sequence* pada persamaan 2.11 dan persamaan 2.12.

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2.11}$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2.12}$$

Sedangkan *feed-forward* network terdiri dari dua transformasi linear dengan fungsi aktivasi ReLU di antaranya. Komponen-komponen ini memastikan model tidak hanya memahami konten tetapi struktur sequential dari dokumen. Urutan informasi seringkali sama pentingnya dengan konten itu sendiri dalam analisis dokumen halal.

Pre-trained Language Models (PLMs) seperti BERT memanfaatkan arsitektur Transformer untuk pembelajaran representasi bahasa yang powerful (Devlin et al., 2018). BERT menggunakan encoder stack dari Transformer yang dilatih pada korpus besar menggunakan masked language modeling dan next sentence prediction. Proses pre-*training* memungkinkan model mempelajari representasi linguistik yang kaya yang dapat di*Transfer* ke berbagai tugas downstream (Devlin et al., 2018). Penelitian memanfaatkan BERT dengan menambahkan Classification Layer di atas representation token [CLS] untuk klasifikasi dokumen halal. Pendekatan ini sangat efektif mengingat terbatasnya data labeled dalam domain halal.

Proses fine-tuning PLMs untuk tugas klasifikasi melibatkan adaptasi model yang telah dilatih sebelumnya ke domain spesifik (Devlin et al., 2018). Secara matematis input *sequence* adalah $x = ([CLS], x_1, \ldots, x_n)$. representasi dari token

[CLS] digunakan untuk klasifikasi pada persamaan 2.13 dan persamaan 2.14 (Vaswani et al., 2017).

$$h = \text{BERT}(x) \tag{2.13}$$

$$P(y \mid x) = \text{softmax}(Wh + b) \tag{2.14}$$

Selama fine-tuning, parameter BERT dan Classification Layer dioptimasi bersama untuk meminimalkan cross-entropy *loss*. Fine-tuning memungkinkan model mengkhususkan diri dalam memahami terminologi dan konteks spesifik sertifikasi halal pada konteks dokumen halal.

Berbagai varian PLMs telah dikembangkan untuk meningkatkan kemampuan pemodelan bahasa. RoBERTa mengoptimalkan proses pre-*training* BERT dengan menghilangkan tugas next sentence prediction dan menggunakan Dynamic masking (Liu et al., 2019). ALBERT memperkenalkan parameter sharing untuk mengurangi konsumsi memori (Lan et al., 2020). DeBERTa mengintegrasikan decoded relative position dan enhanced mask Decoder (P. He et al., 2021). Masing-masing varian ini menawarkan keunggulan berbeda yang dapat dimanfaatkan untuk klasifikasi dokumen halal tergantung pada kebutuhan komputasi dan akurasi.

Integrasi Transformer architecture dengan teknik *Handling ImBalanced* data membuka peluang peningkatan performa klasifikasi. Penelitian mengombinasikan BERT dengan focal *loss* yang memfokuskan pembelajaran pada sampel sulit pada persamaan 2.15.

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{2.15}$$

Alternatif lain dapat menerapkan *Cost-sensitive learning* dengan menyesuaikan *loss function* berdasarkan distribusi kelas. Pendekatan *hybrid* ini memanfaatkan kekuatan representasi BERT sementara secara eksplisit menangani ketidakseimbangan data dokumen halal.

Evaluasi PLMs pada tugas klasifikasi dokumen halal memerlukan pertimbangan metrik yang komprehensif. Metrik seperti *Precision*, *recall*, dan *F1-score* untuk setiap kelas perlu dipantau selain *accuracy*. Fokus terutama untuk kelas

minoritas. Analisis kesalahan dapat mengungkap pola kesalahan sistematis model dalam mengklasifikasikan dokumen halal. Interpretability *Techniques* seperti *attention* visualization dapat membantu memahami keputusan model. Pendekatan evaluasi yang rigor memastikan model tidak hanya akurat tetapi *reliable* dan *trustworthy* untuk aplikasi sertifikasi halal yang kritis.

2.3.3.2 Bidirectional *Encoder Representation*s from *Transformer*s (BERT)

Model BERT memiliki sistem representasi input yang sangat terstruktur. Sistem ini menggabungkan tiga jenis embedding berbeda melalui operasi penjumlahan. Token embeddings mentransformasikan setiap kata menjadi representasi vektor menggunakan WordPiece Tokenizer. Segment embeddings memberikan informasi tentang keanggotaan kalimat untuk setiap token. Position embeddings menginjeksi informasi urutan kata ke dalam model secara eksplisit (Devlin et al., 2018).

WordPiece Tokenizer memproses teks input menjadi unit-unit leksikal. Tokenizer ini membagi kata-kata yang tidak dikenal menjadi subword yang lebih kecil. Token khusus [CLS] selalu menempati posisi pertama dalam setiap *sequence*. Token [SEP] berfungsi sebagai pemisah antara dua kalimat yang berbeda. Vocabulary BERT berisi sekitar 30.000 token yang berbeda untuk berbagai kata dan subword.

Segment embeddings membantu model dalam membedakan antara kalimat pertama dan kedua. Embedding ini memberikan penanda visual kepada model tentang struktur dua kalimat. Semua token dari kalimat A mendapatkan embedding segment yang identik. Semua token dari kalimat B mendapatkan embedding segment yang berbeda. Pembagian ini sangat penting untuk tugas-tugas pemahaman kalimat ganda.

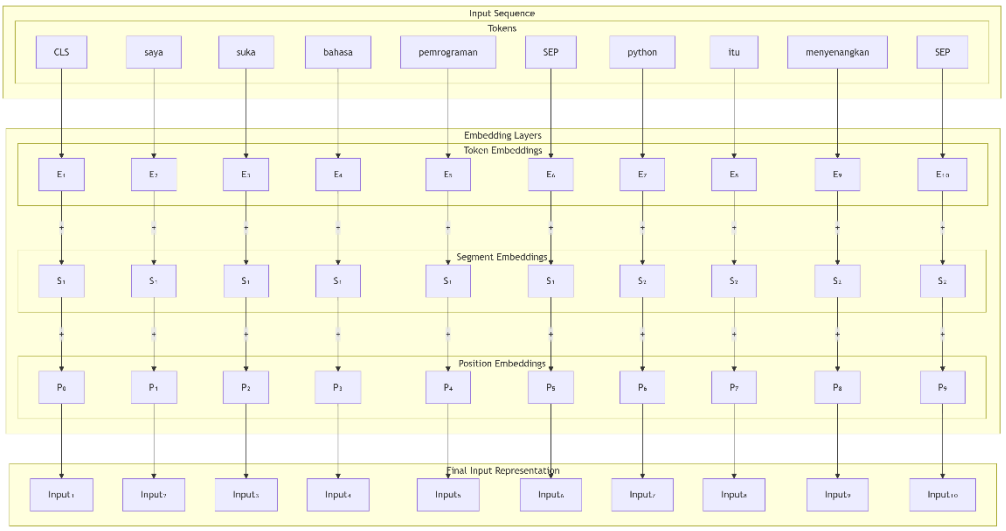
Proses pembentukan input representation mengikuti rumus matematis yang jelas (Devlin et al., 2018). Persamaan 2.16 menjelaskan operasi penjumlahan tersebut. Setiap token dalam *sequence* melalui proses transformasi tiga lapis

embedding ini. Hasil penjumlahan ketiga embedding membentuk input tensor untuk encoder Layers. Proses ini memastikan model menerima informasi yang komprehensif tentang setiap token.

Input = TokenEmbedding + SegmentEmbedding + PositionEmbedding (2.16)

Token [CLS] memegang peran khusus dalam arsitektur BERT. Embedding token ini mengumpulkan informasi kontekstual dari seluruh *sequence*. Model menggunakan representasi token [CLS] untuk tugas klasifikasi teks. Token ini menjadi representasi aggregate untuk keseluruhan input *sequence*. Posisinya yang selalu di awal *sequence* memudahkan model dalam mengakses informasi global. Token [SEP] memiliki fungsi pemisah yang sangat jelas dalam *sequence*. Token ini menandai batas antara dua kalimat yang berbeda. Pemisahan ini membantu model dalam memahami struktur dokumen yang kompleks. Setiap kalimat dalam pasangan mendapatkan segment embedding yang berbeda. Penggunaan token [SEP] konsisten di seluruh arsitektur BERT (Devlin et al., 2018).

Proses tokenization menggunakan WordPiece mampu menangani kata-kata langka dengan efektif. Tokenizer ini memecah kata yang tidak dikenal menjadi subword units. Kemampuan ini meningkatkan cakupan Vocabulary model secara signifikan. Kata-kata teknis dalam dokumen halal dapat diproses dengan lebih akurat. OOV (Out-of-Vocabulary) problems dapat diminimalisir melalui pendekatan ini.



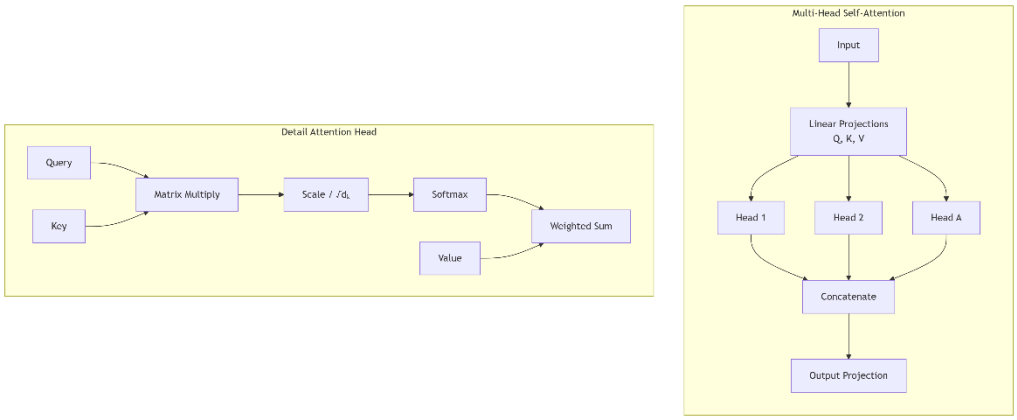
Gambar 2.2 Proses tokenization

BERT membangun representasi inputnya melalui tiga komponen embedding yang dijumlahkan. Model ini mengkonversi setiap kata atau subword menjadi vektor numerik menggunakan Vocabulary WordPiece yang berisi 30.000 token (Devlin et al., 2018). Proses ini selalu diawali dengan token [CLS] untuk tugas klasifikasi dan diakhiri [SEP] sebagai pemisah kalimat. Segment Embeddings kemudian menambahkan informasi keanggotaan kalimat. Semua token dari kalimat pertama mendapat embedding A dan kalimat kedua mendapat embedding B. Pembedaan ini memungkinkan model membedakan konteks antara dua kalimat yang berpasangan. Position Embeddings menginjeksi informasi urutan kata karena Transformer tidak memiliki mekanisme rekurensi seperti RNN. Embedding posisi ini dipelajari untuk setiap posisi dalam urutan hingga panjang maksimum 512 token. Ketiga embedding - token, segment, dan posisi - kemudian dijumlahkan secara element-wise untuk setiap token. Penjumlahan ini menghasilkan vektor input akhir yang kaya akan informasi leksikal, struktural, dan posisional. Representasi gabungan inilah yang menjadi fondasi bagi Layer Transformer BERT untuk melakukan pemrosesan bahasa yang mendalam.

Encoder BERT membentuk inti pemrosesan model melalui tumpukan beberapa Layer Transformer yang identik (Devlin et al., 2018). Arsitektur ini

menggunakan variasi jumlah Layer untuk membedakan kompleksitas model, dimana BERT-Base memiliki 12 Layer dan BERT-*Large* menggunakan 24 Layer. Setiap Layer encoder dalam tumpukan terdiri dari dua sub-Layer utama yang bekerja secara berurutan, yaitu *Multi-head Self-attention* dan *Feed-forward* Neural Network.

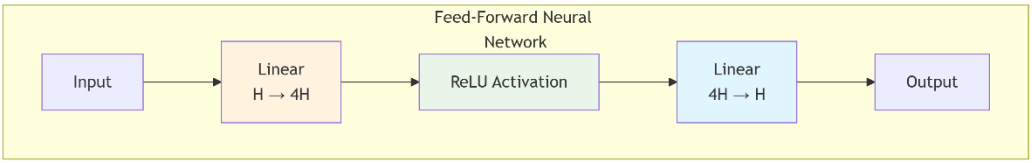
Multi-head Self-attention berfungsi sebagai mekanisme utama untuk memahami konteks antar token secara bidirectional. Mekanisme ini memungkinkan setiap token dalam urutan untuk mempertimbangkan dan memberi bobot pengaruh terhadap semua token lainnya pada dirinya sendiri (Vaswani et al., 2017). BERT mengimplementasikan multiple *attention* heads yang berjalan paralel. Setiap head menggunakan proyeksi Query, Key, dan Value yang berbeda untuk mempelajari berbagai jenis ketergantungan linguistik. Paralelisasi ini memungkinkan model menangkap hubungan yang beragam seperti hubungan subjek-kata kerja dan kata ganti-antecedent dalam representasi yang sama.



Gambar 2.3 *Multi-head Self-attention*

Feed-forward Neural Network berperan sebagai pemroses akhir setelah mekanisme *attention* pada setiap posisi token. Jaringan ini menerapkan transformasi non-linear melalui dua lapisan linear dengan fungsi aktivasi ReLU di antaranya, dengan dimensi 3072 untuk BERT-Base dan 4096 untuk BERT-*Large*. Setiap sub-*Layer* dalam *encoder* dilengkapi dengan koneksi residual dan normalisasi *Layer* untuk menjaga stabilitas pelatihan. Kombinasi antara mekanisme

attention yang memahami konteks global dan FFNN yang melakukan transformasi lokal ini menciptakan representasi yang sangat kaya dan kontekstual untuk pemahaman bahasa.



Gambar 2.4 *Feed-forward* Neural Network

Setiap sub-*Layer* dalam arsitektur BERT menggunakan residual connection dan *Layer* normalization. Residual connection menghubungkan input awal secara langsung ke output sub-*Layer*. *Layer* normalization melakukan stabilisasi distribusi aktivasi di seluruh *Layer*. Kedua mekanisme ini berfungsi mengurangi masalah vanishing *Gradient*. Sistem mempercepat proses konvergensi selama pelatihan model.

Tabel 2.2 Ringkasan Spesifikasi Model

Model	*Layers* (L)	Hidden *Size* (H)	*Attention Heads* (A)	Total Parameters
BERT-Base	12	768	12	~110 Juta
BERT-*Large*	24	1024	16	~340 Juta

Model BERT melalui dua tahap pembelajaran utama (Devlin et al., 2018). Tahap pertama merupakan proses pre-*training* pada data teks tidak berlabel. Tahap kedua merupakan proses *fine-tuning* pada data spesifik berlabel. Kedua tahap ini menggunakan arsitektur model yang identik. Perbedaan utama terletak pada tujuan pembelajaran dan jenis data yang digunakan.

BERT menjalani fase pre-*training* pada data teks berukuran besar. Fase ini

menggunakan dua tugas unsupervised sebagai tujuan pembelajaran. Masked Language Model (MLM) menyembunyikan token input secara acak. Next Sentence Prediction (NSP) melatih model memahami hubungan antar kalimat. MLM membuat model memprediksi token yang disembunyikan berdasarkan konteksnya. NSP menentukan apakah suatu kalimat merupakan kelanjutan logis dari kalimat sebelumnya.B. *Fine-tuning* (Fase Penyesuaian).

Proses *fine-tuning* mengadaptasi model BERT yang telah dilatih sebelumnya untuk tugas-tugas spesifik (Devlin et al., 2018). Penambahan satu lapisan output sederhana melengkapi arsitektur dasar model. Lapisan tambahan berfungsi sebagai klasifier untuk tugas tertentu. Optimasi seluruh parameter model berjalan bersamaan pada *dataset* baru. Pendekatan ini memanfaatkan pengetahuan linguistik dari proses pre-*training*. Metode menghasilkan kinerja optimal dengan data pelatihan terbatas.

Model BERT men*Gap*likasikan pendekatan berbeda untuk berbagai tugas pemrosesan bahasa. Tugas klasifikasi kalimat memanfaatkan output token [CLS] sebagai representasi seluruh input untuk klasifikasi. Sistem memproses output token [CLS] melalui *feed-forward* neural network dan fungsi softmax. Tugas klasifikasi token menggunakan output setiap token individual melalui *Classifier* terpisah. Proses question-answer memerlukan lapisan linear tambahan untuk memprediksi token awal dan akhir jawaban. Tugas paraphrase memanfaatkan kembali output token [CLS] untuk menentukan kemiripan semantik antara dua kalimat.

2.3.3.3 IndoBERT: BERT untuk Bahasa Indonesia

IndoBERT merupakan model bahasa pre-trained BERT yang secara khusus dikembangkan untuk bahasa Indonesia (Koto et al., 2021). Penelitian dan pengembangan model ini dilakukan melalui kolaborasi antara Universitas AI Indonesia (UAi) dan tim IndoNLU. Model ini memiliki beberapa perbedaan mendasar dengan BERT versi Original dalam hal implementasinya (Wilie et al., 2020). Proses pre-*training* menggunakan kumpulan data teks berbahasa Indonesia

dalam skala sangat besar. Vocabulary Tokenizer dioptimalkan secara khusus untuk morfologi bahasa Indonesia yang kompleks. Performa model menunjukkan keunggulan signifikan dibandingkan BERT multilingual dalam berbagai tugas NLP.

Proses pre-*training* IndoBERT memanfaatkan kumpulan data teks berbahasa Indonesia yang ekstensif dan beragam (Cahyawijaya et al., 2021). Data *training* mencakup korpus teks dari berbagai domain termasuk berita, sastra, dan percakapan sehari-hari. Tokenizer model menggunakan Vocabulary yang disesuaikan dengan karakteristik morfologis bahasa Indonesia (Koto et al., 2020). Penyesuaian Vocabulary ini mampu menangani prefiks, sufiks, dan konfiks yang umum dalam pembentukan kata bahasa Indonesia. Optimasi Tokenizer meningkatkan kemampuan model dalam memahami kata-kata turunan dan bentuk dasar. Model dapat menangani variasi morfologis kata dengan lebih efektif.

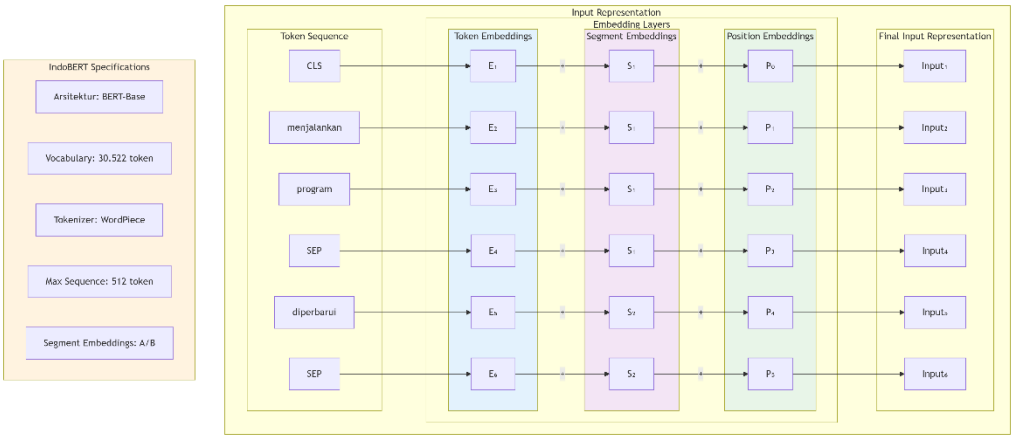
Arsitektur IndoBERT mengadopsi desain BERT-Base tanpa modifikasi signifikan pada struktur jaringan (Koto et al., 2020). Model tetap mempertahankan konfigurasi 12 Layer encoder dengan 768 hidden dimensions dan 12 *attention* heads. Inovasi utama terletak pada proses pre-*training* dan tokenization yang disesuaikan dengan karakteristik bahasa Indonesia (Wilie et al., 2020). Pendekatan ini menghasilkan performa yang mengungguli BERT multilingual dalam berbagai benchmark NLP bahasa Indonesia. Keunggulan mencakup tugas klasifikasi teks, named entity recognition, dan question answering (Cahyawijaya et al., 2021). Fokus spesifik pada bahasa Indonesia membuat IndoBERT menjadi pilihan optimal untuk aplikasi NLP domestik.

IndoBERT secara fundamental dibangun di atas arsitektur BERT-Base yang telah teruji. Model ini menggunakan WordPiece Tokenizer dengan Vocabulary berisi 30.522 token yang dipelajari secara statistik dari korpus bahasa Indonesia yang sangat besar. Vocabulary yang khusus ini memungkinkan Tokenizer memahami secara efektif struktur subword dan kata-kata yang umum dalam bahasa Indonesia seperti "menjalankan", "diperbarui", dan "perlengkapan".

Sistem embedding IndoBERT terdiri dari tiga komponen utama yang

dijumlahkan untuk merepresentasikan input. Token Embeddings secara khusus dioptimalkan untuk kosa kata bahasa Indonesia, sementara Segment Embeddings berfungsi membedakan antara kalimat pertama (Segment A) dan kalimat kedua (Segment B) dalam pasangan teks. Segment Embeddings ini sangat berguna untuk tugas-tugas seperti Natural Language Inference dan Question Answering yang melibatkan sepasang teks.

Position Embeddings melengkapi sistem dengan memberikan informasi urutan token dalam teks. Model mempelajari embedding untuk setiap posisi token hingga panjang 512 token, sama seperti BERT asli. Ketiga komponen embedding kemudian diintegrasikan melalui penjumlahan langsung TokenEmbedding + SegmentEmbedding + PositionEmbedding. Representasi input yang dihasilkan ini teroptimasi untuk karakteristik linguistik bahasa Indonesia sekaligus mempertahankan kemampuan arsitektur BERT yang andal.



Gambar 2.5 Position *Embedding*s

Encoder stack merupakan inti model IndoBERT yang terdiri dari 12 Layer encoder identik. Setiap Layer encoder memiliki dimensi tersembunyi sebesar 768 dimensi dengan 12 *attention* head yang bekerja paralel. Arsitektur ini mengadopsi desain BERT-Base secara utuh tanpa modifikasi signifikan. *Multi-head self-attention* mechanism memungkinkan setiap token berinteraksi dengan semua token lainnya dalam *sequence*. Position-wise *feed-forward* network memproses

representasi token secara independen pada setiap posisi. Residual connection dan Layer normalization menstabilkan proses *training* di seluruh Layer yang dalam.

Multi-head self-attention mechanism membagi perhitungan *attention* menjadi 12 head yang beroperasi secara paralel. Setiap head menghitung perhatian dengan dimensi 64 yang berasal dari pembagian 768 dimensi dengan 12 head. Mekanisme ini memungkinkan model menangkap berbagai jenis ketergantungan linguistik dari perspektif yang berbeda. Setiap head mempelajari pola-pola perhatian yang kompleks dan saling melengkapi. Formula *attention* yang digunakan sama persis dengan arsitektur Transformer Original. Kemampuan bidirectional dari mekanisme ini menjadi kunci pemahaman konteks utuh dalam bahasa Indonesia.

Position-wise *feed-forward* network menerapkan transformasi non-linear pada setiap posisi token secara independen. Jaringan ini terdiri dari dua lapisan linear dengan fungsi aktivasi Gaussian Error Linear Unit di antaranya. Lapisan pertama memperluas dimensi dari 768 menjadi 3072 untuk meningkatkan kapasitas model. Lapisan kedua mengembalikan dimensi dari 3072 menjadi 768 untuk konsistensi dimensi output. Residual connection menghubungkan input langsung ke output setiap sub-Layer melalu operasi penjumlahan. Layer normalization menstabilkan distribusi aktivasi dan mempercepat konvergensi *training* model.

Tabel 2.3 Ringkasan Spesifikasi Model IndoBERT

Parameter	Nilai untuk IndoBERT	Keterangan
Arsitektur Dasar	BERT-Base	-
Jumlah *Layer* (L)	12	Sama dengan BERT-Base
Dimensi Tersembunyi (H)	768	Sama dengan BERT-Base

Parameter	Nilai untuk IndoBERT	Keterangan
Jumlah *Attention Head* (A)	12	Sama dengan BERT-Base
Dimensi FFNN	4*H = 3072	Sama dengan BERT-Base
Panjang Maksimal Urutan	512 token	Sama dengan BERT-Base
Vocabulary Size	30,522	Diolah dari korpus Bahasa Indonesia
Total Parameters	~124 Juta	Sedikit lebih banyak dari BERT-Base (~110J) karena *Vocabulary* yang sedikit lebih besar

Proses pre-*training* menjadi faktor kunci keberhasilan IndoBERT yang membedakannya dari model multibahasa. Pelatihan awal ini secara khusus dirancang untuk karakteristik bahasa Indonesia. IndoBERT menjalani pre-*training* pada Indonesian Cascade Corpus yang sangat besar dan terkurasi. Korpus ini mengumpulkan teks dari berbagai sumber terpercaya seperti Wikipedia Indonesia dan portal berita ternama. Proses kurasi ketat memastikan kualitas data *training* yang konsisten dan representatif. Spesifisitas inilah yang memberikan keunggulan performa dibandingkan model multibahasa.

Indonesian Cascade Corpus mencakup lebih dari 200 juta kalimat dengan total 3.5 miliar kata. Koleksi data yang masif ini berasal dari domain beragam termasuk media online, literatur, dan konten web umum. Keragaman domain memungkinkan model mempelajari variasi gaya bahasa dan kosakata. Dua tugas pre-*training* diterapkan secara konsisten yaitu Masked Language Modeling dan

Next Sentence Prediction. Masked Language Modeling melatih model memahami konteks secara bidirectional dengan memprediksi token yang disembunyikan. Next Sentence Prediction mengembangkan kemampuan memahami hubungan logis antar kalimat.

Sistem tokenisasi IndoBERT mengoptimalkan Vocabulary WordPiece berdasarkan korpus bahasa Indonesia. Optimasi ini menghasilkan kemampuan menangani morfologi bahasa yang kaya prefiks dan sufiks. Kata turunan seperti "berjalan" dan "perjalanan" terpecah menjadi subword meaningful "ber##jalan" dan "per##jalan##an". Model efektif memproses kata serapan dari bahasa asing dengan akurasi tinggi. Efisiensi tokenisasi mengurangi jumlah token unknown dalam pemrosesan teks. Pendekatan spesifik-domain ini meningkatkan performa model pada berbagai tugas NLP bahasa Indonesia.

Proses fine-tuning mengadaptasi model IndoBERT untuk tugas-tugas spesifik. Penambahan satu lapisan output menyederhanakan proses adaptasi ini. Tugas klasifikasi kalimat memanfaatkan output token [CLS]. Klasifikasi token menggunakan output setiap token individu. Sistem pertanyaan-jawaban memprediksi token awal dan akhir jawaban. Tugas kesamaan teks membandingkan output [CLS] dari pasangan kalimat.

2.3.4 Analisis Distribusi Data pada Domain Sertifikasi Halal

Penelitian ini mengungkap pola ketidakseimbangan alami pada *dataset* halal. Popularitas produk makanan mendominasi *dataset* dibandingkan produk farmasi. Tingkat kompleksitas produk mempengaruhi distribusi data secara signifikan. Regulasi daerah yang berbeda menciptakan variasi distribusi berdasarkan region. Faktor-faktor ini membentuk pola distribusi yang unik dalam domain halal.

Distribusi data dokumen halal mengikuti pola power-law pada jenis produk. Sejumlah kecil kategori produk mendominasi sebagian besar *dataset*. Distribusi multi-modal muncul pada kategori kompleksitas produk. Hierarchical ImBalance terjadi pada level klasifikasi yang berbeda. Pola ketidakseimbangan bertingkat

mempengaruhi kinerja model klasifikasi. Karakteristik distribusi ini memerlukan pendekatan penanganan yang khusus.

Analisis distribusi memberikan dasar untuk pengembangan teknik *sampling* yang efektif. Pemahaman pola distribusi membantu memilih metode pre*processing* yang tepat. Karakteristik hierarchical ImBalance membutuhkan pendekatan bertingkat. Variasi berdasarkan region mempengaruhi generalisasi model klasifikasi. Pola multi-modal memerlukan teknik *Handling* yang adaptif. Penanganan yang tepat meningkatkan kinerja model pada semua kategori produk.

2.3.4 Teori *Sampling* dan Distribusi Data

2.3.4.1 *Statistical* Foundations of *Sampling*

Penny (1995) menetapkan fondasi teori *sampling* dalam penelitian statistik modern. Prinsip representativitas menjadi landasan utama dalam pemilihan sampel penelitian. Sampel penelitian harus merepresentasikan distribusi populasi secara akurat dan komprehensif. *Sampling* bias dapat menyebabkan model statistik yang bias dan tidak valid. Estimasi parameter populasi mengikuti rumus dasar pada persamaan 2.17. Teori *sampling* menjamin generalisasi hasil penelitian ke populasi target (Goodfellow et al., 2014). Jenis-jenis teknik *sampling* terbagi dalam dua kategori utama yang berbeda. *Probability sampling* menggunakan pendekatan acak berbasis probabilitas. Non-*Probability sampling* mengandalkan pertimbangan subjektif peneliti. Pemilihan teknik *sampling* mempengaruhi validitas eksternal penelitian.

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.17}$$

Probability sampling menggunakan metode pemilihan sampel secara acak dan sistematis. Simple random *sampling* memilih sampel dengan peluang yang sama untuk semua unit. Probabilitas pemilihan setiap unit dalam Simple random *sampling* pada persamaan 2.18. Stratified *sampling* membagi populasi ke dalam strata homogen berdasarkan karakteristik. Cluster *sampling* memilih sampel berdasarkan kelompok alami yang sudah terbentuk. Systematic *sampling* memilih

sampel dengan interval tetap dari kerangka *sampling*. Metode *Probability sampling* menjamin representativitas sampel melalui prinsip acak. Teknik ini mengurangi bias seleksi dalam proses pengambilan sampel (J. M. Johnson & Khoshgoftaar, 2019). *Probability sampling* memungkinkan perhitungan error *sampling* secara kuantitatif. Presisi estimasi dapat ditingkatkan melalui desain *sampling* yang tepat.

$$\mathrm{P} = \frac{1}{N} \tag{2.18}$$

Stratified *sampling* meningkatkan efisiensi estimasi parameter populasi secara signifikan (Goodfellow et al., 2014). Varians estimasi mean strata dihitung dengan persamaan 2.19. Alokasi sampel proporsional menggunakan pada persamaan 2.20. Cluster *sampling* menggunakan rumus efisiensi relative pada pesamaan 2.21. Simple random *sampling* memiliki varians estimasi pada persamaan 2.22. Two-stage *sampling* memerlukan perhitungan varians yang lebih kompleks. Optimal allocation dalam stratified *sampling* meminimalkan varians total. Desain *sampling* mempertimbangkan *Trade-off* antara biaya dan presisi. Perencanaan *sampling* yang baik mempertimbangkan heterogenitas populasi. Analisis statistik memerlukan penyesuaian untuk desain *sampling* kompleks.

$$\sigma_{\text{str}}^2 = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} \tag{2.19}$$

$$n_h = n \times \frac{N_h}{N} \tag{2.20}$$

$$RE = 1 + (m-1)\rho \tag{2.21}$$

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \tag{2.22}$$

Non-*Probability sampling* tidak menggunakan prinsip acak dalam seleksi sampel penelitian. Convenience *sampling* memilih sampel berdasarkan kemudahan akses dan ketersediaan (Fernández et al., 2018). Purposive *sampling* memilih sampel dengan kriteria khusus yang telah ditetapkan. Quota *sampling* menentukan kuota berdasarkan proporsi karakteristik populasi. Snowball *sampling* mengandalkan referensi dari responden yang sudah terpilih. Estimasi bias dalam Non-*Probability sampling* diukur dengan persamaan 2.23. Teknik ini cocok untuk populasi yang sulit diakses atau tersembunyi. Non-*Probability sampling* memiliki

keterbatasan dalam generalisasi hasil penelitian. Metode ini sering digunakan dalam penelitian kualitatif dan eksploratori. Validitas hasil sangat bergantung pada expertise peneliti.

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{2.23}$$

Prinsip representativitas mensyaratkan akurasi representasi sampel terhadap populasi. Margin of error dihitung dengan persamaan 2.24. *Sampling* frame harus mencakup seluruh unit dalam populasi target. *Coverage* error terjadi ketika kerangka *sampling* tidak lengkap. Non-response bias mempengaruhi representativitas sampel akhir. Measurement error berasal dari instrumen pengumpulan data yang tidak valid (Fernández et al., 2018). *Sampling Weight* mengkompensasi probabilitas seleksi yang tidak equal. Post-stratification *adjustment* memperbaiki ketidakrepresentatifan sampel. Quality control memastikan proses *sampling* sesuai protokol. Audit *sampling* memverifikasi akurasi proses seleksi sampel.

$$ME = z\sqrt{\frac{p(1-p)}{n}} \tag{2.24}$$

Ukuran sampel minimum ditentukan oleh persamaan 2.25. Finite population correction factor pada persamaan 2.26. Power analysis menggunakan persamaan 2.27. Convenience *sampling* mengutamakan efisiensi waktu dan biaya penelitian (Fernández et al., 2018). Purposive *sampling* cocok untuk studi kasus spesifik dan penelitian kualitatif. Maximum variation *sampling* menangkap keragaman fenomenon penelitian. Theoretical *sampling* dalam grounded theory mengembangkan teori emergent. Non-*Probability sampling* memerlukan kehati-hatian dalam interpretasi hasil. Mixed-methods *sampling* mengintegrasikan pendekatan kuantitatif dan kualitatif. Pemahaman fondasi statistik *sampling* menjamin kualitas dan keandalan penelitian.

$$n = \frac{Z^2\sigma^2}{E^2} \tag{2.25}$$

$$fpc = \sqrt{\frac{N-n}{N-1}} \tag{2.26}$$

$$1 - \beta = \Phi(z - z_{1-\alpha/2}) \quad (2.27)$$

2.3.4.2 *Sampling Theory* for *ImBalanced* Data

Fernández et al. (2018) mengembangkan teori *sampling* khusus untuk data tidak seimbang. Penelitian merumuskan strategic *sampling framework* yang komprehensif. *Framework* ini mengatasi masalah ketidakseimbangan distribusi kelas dalam *dataset*. Strategic *sampling framework* terdiri dari dua pendekatan utama yang saling melengkapi. Pendekatan pertama adalah informed *undersampling* untuk kelas mayoritas. Pendekatan kedua adalah strategic *oversampling* untuk kelas minoritas. Kedua metode ini bekerja sinergis untuk menyeimbangkan distribusi data. Implementasi *framework* ini memerlukan pemahaman karakteristik *dataset*. Teori ini memberikan landasan matematis untuk teknik *sampling* adaptif. Aplikasi *framework* meningkatkan kinerja model klasifikasi pada kelas minoritas.

Informed *undersampling* secara selektif menghapus sampel mayoritas yang redundant. Proses ini mempertahankan sampel informatif yang penting untuk *decision boundary* (Fernández et al., 2018). Redundancy measure dihitung menggunakan pesamaan 2.28. Sample removal *Probability* berbanding lurus dengan nilai Redundancy. Kriteria informativess menggunakan entropy pada persamaan 2.29. Proses *undersampling* mempertahankan structural integrity *dataset*. Metode ini mencegah hilangnya informasi penting dari kelas mayoritas. Optimasi *undersampling* mencapai *Trade-off* antara balancing dan informasi. Implementasi menggunakan nearest neighbor analysis untuk identifikasi Redundancy. Hasil akhir adalah *dataset* mayoritas yang condensed namun informatif.

$$R(x) = \frac{1}{k}\sum_{i=1}^{k} d(x, \mathrm{NN}_i(x)) \quad (2.28)$$

$$E(x) = -\sum p(c \mid x) \log p(c \mid x) \quad (2.29)$$

Strategic *oversampling* menambah sampel minoritas di daerah *decision boundary* (Fernández et al., 2018). Synthetic sample generation menggunakan interpolasi pada persamaan 2.30. Parameter λ mengontrol diversitas sampel sintetik

yang dihasilkan. *Boundary* detection menggunakan margin theory pada persamaan 2.31. *Probability sampling Weight* pada persamaan 2.32. Strategic placement mencegah *overfitting* dengan menjaga diversitas sampel. Density estimation memastikan distribusi sampel sesuai dengan underlying *distribution*. Adaptive kernel bandwidth pada persamaan 2.33. *Oversampling* factor ditentukan berdasarkan ImBalance ratio. Metode ini menghasilkan sampel minoritas yang meaningful dan diverse.

$$x_{\text{new}} = x_i + \lambda(x_j - x_i) \tag{2.30}$$

$$M(x) = \frac{|f(x)|}{\|\nabla f(x)\|} \tag{2.31}$$

$$w(x) = \exp\left(-\frac{M(x)}{\sigma}\right) \tag{2.32}$$

$$\sigma = \text{median}(\| x_i - x_j \|) \tag{2.33}$$

Integrasi informed *undersampling* dan strategic *oversampling* menciptakan equilibrium (Fernández et al., 2018). *Combined sampling* ratio dengan persamaan 2.34. Optimal Balance dicapai ketika $\alpha \approx 1$ dengan Constraint quality. Quality metric pada persamaan 2.35. *Framework* menggunakan Adaptive stopping criterion berdasarkan *Convergence* measure. *Convergence* diukur dengan persamaan 2.36. Implementasi iteratif menyesuaikan *sampling* parameters secara dinamis. Monitoring menggunakan *learning curves* untuk prevent Over-*sampling*. Hasil integrasi adalah *dataset* yang seimbang dan representatif. Validasi menggunakan cross-*Validation* dengan stratified *sampling*.

$$\alpha = \frac{n_{\text{minority,new}}}{n_{majority,\text{new}}} \tag{2.34}$$

$$Q = \frac{1}{2}[D_{\text{KL}}(p \,\|\, \hat{p}) + D_{\text{KL}}(q \,\|\, \hat{q})] \tag{2.35}$$

$$\Delta Q = | Q_t - Q_{t-1} | < \varepsilon \tag{2.36}$$

Strategic *sampling framework* mengoptimalkan multiple *Objectives* secara simultan (Fernández et al., 2018). *Objective function* pada persamaan 2.37. *Regularization* term mencegah *overfitting* pada persamaan 2.38. Multi-*Objective Optimization* menggunakan *Pareto optimality concept*. *Weight* adaptation pada persamaan 2.39. *Framework incorporates* domain *Knowledge* melalui *Constraints*.

Constrained Optimization pada min J(θ) subject to $g_j(\theta) \leq 0$. *Hyperparameter* tuning menggunakan Bayesian *Optimization*. Model selection berdasarkan *Balanced accuracy* metric. Evaluasi komprehensif mencakup *Statistical significance testing*.

$$J(\theta) = \alpha_1 \, Accuracy(\theta) + \alpha_2 \, \mathrm{F1}_{\mathrm{minority}}(\theta) + \alpha_3 \, Diversity(\theta) \quad (2.37)$$

$$R(\theta) = \beta \parallel \theta \parallel^2 \quad (2.38)$$

$$\alpha_i = \mathrm{softmax}(z_i/\tau) \quad (2.39)$$

2.3.4.3 *Distribution*-Aware *Sampling*

J. Johnson et al. (2019) memperkenalkan konsep *distribution*-aware *sampling* dalam penelitian terbaru mereka. Pendekatan ini mempertimbangkan distribusi data secara eksplisit dalam proses *sampling*. *Density-based sampling* menjadi teknik inti dalam *framework* yang mereka usulkan. Probabilitas pemilihan sampel berbanding terbalik dengan kepadatan data di sekitarnya. Rumus matematisnya dinyatakan sebagai persamaan 2.40. Teknik ini lebih memilih sampel yang berada di daerah *low-density regions*. Tujuannya adalah meningkatkan *Coverage* ruang fitur secara keseluruhan. *Distribution*-aware *sampling* mengatasi masalah *under-representation* area *sparse* (Fernández et al., 2018). Metode ini efektif untuk *dataset* dengan distribusi yang kompleks. Implementasinya memerlukan estimasi density *function* yang akurat.

$$p(x) \propto \frac{1}{density(x)} \quad (2.40)$$

Density-based sampling menggunakan kernel density estimation untuk mengukur kepadatan (Fernández et al., 2018). Estimasi density *function* dihitung dengan persamaan 2.41. Fungsi kernel K biasanya menggunakan Gaussian kernel persamaan 2.42. Bandwidth parameter h dioptimalkan menggunakan Silverman's rule pada persamaan 2.43. *Probability* selection *Weight* menggunakan inverse density pada persamaan 2.44. *Regularization* term ε mencegah *division by zero*. Proses *sampling* menggunakan *Weight*ed random *sampling* berdasarkan *Weight* w(x). Hasilnya adalah representasi yang lebih baik untuk *sparse regions*. Teknik ini

meningkatkan *Diversity* sampel *training*. Optimasi bandwidth crucial untuk Performance density estimation.

$$density(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (2.41)$$

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right) \quad (2.42)$$

$$h = 1.06\,\sigma\, n^{-1/5} \quad (2.43)$$

$$w(x) = \frac{1}{density(x)+\varepsilon} \quad (2.44)$$

Low-density regions sampling meningkatkan *Coverage* ruang fitur secara signifikan (Fernández et al., 2018). *Coverage* metric diukur dengan persamaan 2.45. Feature space partitioning menggunakan Voronoi tessellation untuk analisis *Coverage*. *Boundary* samples diidentifikasi menggunakan margin-based criterion. Margin dihitung dengan persamaan 2.46. *Sampling* di *decision boundary* menggunakan *Probability*. *Temperature parameter* T mengontrol *exploration-exploitation Trade-off*. Strategic *sampling* di *boundary* meningkatkan model *discriminative* power. Metode ini efektif untuk improving model *decision boundary*. Hasilnya adalah peningkatan kinerja klasifikasi secara keseluruhan.

$$C = \frac{1}{m}\sum_{j=1}^{m} \exp(-\min_i d(x_j, x_i)) \quad (2.45)$$

$$M(x) = \frac{|f(x)|}{\|\nabla f(x)\|} \quad (2.46)$$

Distribution-aware *sampling* mengintegrasikan multiple criteria dalam *framework* terpadu (Fernández et al., 2018). *Combined Weight*ting *function* dengan persamaan 2.47. *Diversity Weight* menggunakan *repulsiveness measure* pada persamaan 2.48. *Optimization Objective* memaksimalkan pada persamaan 2.49. Adaptive parameter tuning menggunakan *Gradient*-based *Optimization*. Constraint *Handling* memastikan *feasibility solution space*. Multi-criteria *decision making* menggunakan *Pareto Optimization*. *Framework Supports different* types of *distribution*s dan patterns. Implementasi menggunakan *efficient algorithm* untuk *Large dataset*s. *Validation* menggunakan *comprehensive evaluation metrics*.

$$w_{\text{total}}(x) = w_{density}(x) \times w_{boundary}(x) \times w_{Diversity}(x) \quad (2.47)$$

$$w_{Diversity}(x) = \min_i d(x, x_i^{\text{selected}}) \quad (2.48)$$

$$J = \alpha\ Coverage\ +\ \beta\ Diversity\ +\ \gamma\ Boundary\_quality \quad (2.49)$$

Aplikasi *distribution*-aware *sampling* memerlukan validasi empiris yang komprehensif (Fernández et al., 2018). Experimental design menggunakan multiple benchmark *dataset*s. Performance *evaluation* menggunakan *metrics*: *Balanced Accuracy*, *F1-score*, G-mean. *Statistical testing* menggunakan Wilcoxon signed-rank test. Effect Size diukur menggunakan *Rank-biserial correlation*. Comparative analysis terhadap state-of-the-art methods. Ablation study mengisolasi contribution setiap komponen. Sensitivity analysis terhadap *Hyperparameter*s *framework*. *Robustness evaluation* across *different* ImBalance *ratios*. Hasil eksperimen menunjukkan superioritas *distribution*-aware *sampling*.

2.3.5 Teknik *Sampling* untuk Penanganan *ImBalance*

2.3.5.1 *Oversampling Techniques*

Chawla et al. (2002) melakukan kategorisasi komprehensif terhadap teknik *oversampling* dalam pembelajaran mesin. Kategorisasi ini membagi teknik *oversampling* menjadi beberapa pendekatan fundamental. Random *oversampling* merupakan metode paling sederhana dalam kategori ini (Fernández et al., 2018). *Synthetic Minority Over-sampling Technique* (SMOTE) merepresentasikan pendekatan yang lebih canggih. He, H., Bai, Y., Garcia, E., & Li (2008) mengembangkan ADASYN sebagai penyempurnaan dari SMOTE. Setiap teknik memiliki karakteristik dan aplikasi yang berbeda-beda. Pemilihan teknik *oversampling* bergantung pada karakteristik *dataset*. Implementasi yang tepat meningkatkan kinerja model klasifikasi. *Oversampling* efektif untuk menangani ketidakseimbangan kelas minoritas. Evaluasi performa menentukan efektivitas teknik yang digunakan.

Random *oversampling* melakukan duplikasi sampel minoritas secara acak (Fernández et al., 2018). Proses ini meningkatkan jumlah sampel kelas minoritas hingga seimbang. Probabilitas duplikasi setiap sampel adalah persamaan 2.50. Jumlah duplikasi untuk setiap sampel mengikuti distribusi seragam. Total sampel

baru dihitung dengan persamaan 2.51. Metode ini sederhana dalam implementasi dan komputasi. Namun, random *oversampling* rentan terhadap *overfitting*. Duplikasi sampel yang sama mengurangi diversitas data. Model menjadi kurang *robust* dalam generalisasi. Teknik ini cocok untuk *dataset* dengan variasi terbatas.

$$P(x) = \frac{1}{n_{\text{minority}}} \tag{2.50}$$

$$n_{\text{new}} = n_{majority} - n_{\text{minority}} \tag{2.51}$$

SMOTE menghasilkan sampel sintetik melalui interpolasi linier (Chawla et al., 2002). Algoritma memilih sampel minoritas x_i dan tetangganya x_{z_i}. Sampel sintetik dibentuk dengan persamaan 2.52. Parameter λ merupakan bilangan random antara 0 dan 1. Pemilihan tetangga menggunakan k-NN dengan metric Euclidean. Jumlah tetangga k biasanya bernilai 5. Variasi SMOTE menggunakan *different* neighbor selection strategies. Borderline-SMOTE fokus pada sampel di *decision boundary* (Fernández et al., 2018). SVM-SMOTE menggunakan *Support vectors* untuk generasi sampel. Adaptive-SMOTE menyesuaikan parameter berdasarkan local density.

$$x_{\text{new}} = x_i + \lambda(x_{z_i} - x_i) \tag{2.52}$$

ADASYN mengembangkan SMOTE dengan pendekatan adaptif (He, H., Bai, Y., Garcia, E., & Li, 2008). Teknik ini menyesuaikan distribusi sampel sintetik berdasarkan kesulitan *learning*. Density *distribution* diestimasi dengan persamaan 2.53. Kesulitan *learning* diukur menggunakan rasio misklasifikasi. *Weight distribution* dihitung dengan persamaan 2.54. Jumlah sampel sintetik per titik pada persamaan 2.55. Total sampel yang dihasilkan persamaan 2.56. Parameter β mengontrol derajat keseimbangan yang diinginkan. ADASYN lebih fokus pada daerah yang sulit dipelajari. Hasilnya adalah *improvement* signifikan pada *decision boundary*.

$$d_i = \frac{\Delta_i}{K} \tag{2.53}$$

$$w_i = \frac{d_i}{\sum_j d_j} \tag{2.54}$$

$$g_i = w_i \times n_{Synthetic} \quad (2.55)$$

$$n_{Synthetic} = \beta \times (n_{majority} - n_{\text{minority}}) \quad (2.56)$$

Implementasi teknik *oversampling* mempertimbangkan karakteristik ruang fitur. Feature space analysis menggunakan PCA untuk visualisasi (He, H., Bai, Y., Garcia, E., & Li, 2008). Distance metric selection mempengaruhi kualitas sampel sintetik. Euclidean distance cocok untuk Continuous Features. Manhattan distance lebih *robust* terhadap *outlier*s. Mahalanobis distance mempertimbangkan korelasi fitur. Pre*processing* yang tepat meningkatkan efektivitas *oversampling*. Normalisasi fitur mencegah dominasi fitur skala besar. *Handling* categorical Features memerlukan pendekatan khusus. Mixed-type data membutuhkan distance metric yang *hybrid*.

Evaluasi teknik *oversampling* menggunakan multiple Performance *metrics* (Fernández et al., 2018). *Balanced accuracy* pada persamaan 2.57. Geometric mean pada persamaan 2.58. *F1-score* pada persamaan 2.59. AUC-ROC mengukur *Overall* Classification Performance. *Statistical significance* diuji dengan paired t-test. Cross-*Validation* mencegah *overfitting* dalam evaluasi. Perbandingan dilakukan terhadap baseline tanpa *sampling*. Analysis of variance menguji perbedaan antar teknik. Runtime complexity menjadi pertimbangan praktis. Scalability terhadap *Large dataset*s perlu diperhitungkan.

$$\text{BA} = \frac{1}{2}(\text{TPR} + \text{TNR}) \quad (2.57)$$

$$\text{G-mean} = \sqrt{\text{TPR} \times \text{TNR}} \quad (2.58)$$

$$\text{F1} = 2 \times \frac{Precision \times recall}{Precision + recall} \quad (2.59)$$

Pemilihan teknik *oversampling* optimal memerlukan pertimbangan menyeluruh. *Dataset* characteristics menentukan suitability teknik. Computational resources membatasi pilihan implementasi. Domain *Knowledge* memandu parameter tuning. *Ensemble* methods mengkombinasikan multiple teknik. *Hybrid* approaches mengintegrasikan *oversampling* dan *undersampling*. Dynamic selection menyesuaikan teknik selama *training*. Automated *machine learning* menyederhanakan proses seleksi. Future research mengembangkan teknik yang

lebih adaptif. Integration dengan deep *learning* architectures menjadi tren terkini.

2.3.5.2 *Adaptive Synthetic Sampling* (ADSYN)

He, H., Bai, Y., Garcia, E., & Li (2008) mengembangkan *Adaptive Synthetic Sampling* (ADASYN) sebagai teknik *oversampling* yang canggih. ADASYN mengatasi keterbatasan SMOTE dengan pendekatan yang adaptif dan terarah. Teknik ini secara otomatis menyesuaikan jumlah sampel sintetik yang dihasilkan untuk setiap titik data minoritas. Penyesuaian didasarkan pada tingkat kesulitan *learning* setiap instance. Instance yang lebih sulit dipelajari menerima lebih banyak sampel sintetik. Adaptive *Weight* calculation menggunakan persamaan 2.60 . Kesulitan *learning* diestimasi menggunakan rasio misklasifikasi lokal. *Distribution*-sensitive allocation menghitung jumlah sampel per titik pada persamaan 2.61. Pendekatan ini meningkatkan fokus pada daerah *decision boundary* yang kompleks. Hasilnya adalah peningkatan kinerja model untuk kelas minoritas.

$$w_i = \frac{\text{difficulty}_i}{\sum_{j=1}^{m} \text{difficulty}_j} \tag{2.60}$$

$$n_i = w_i \times N_{\text{total}} \tag{2.61}$$

Proses ADASYN dimulai dengan menghitung tingkat kesulitan *learning* untuk setiap sampel minoritas (He, H., Bai, Y., Garcia, E., & Li, 2008). Tingkat kesulitan diukur berdasarkan proporsi sampel mayoritas di antara k-tetangga terdekat. Rumus difficulty pada persamaan 2.62. Parameter k biasanya bernilai 5 dalam implementasi standar. Normalisasi difficulty values menghasilkan *Weight distribution* yang valid. Jumlah total sampel sintetik ditentukan pada persamaan 2.63. Parameter β mengontrol derajat keseimbangan yang diinginkan. Alokasi sampel per titik memastikan distribusi yang proporsional. Generasi sampel sintetik menggunakan interpolasi linier seperti SMOTE. Implementasi yang efisien memanfaatkan struktur data k-d tree.

$$\text{difficulty}_i = \frac{\text{number of } \textit{majority} \text{ neighbors}}{k} \tag{2.62}$$

$$N_{\text{total}} = \beta \times (n_{\textit{majority}} - n_{\text{minority}}) \tag{2.63}$$

Evaluasi ADASYN menunjukkan keunggulan signifikan dibandingkan teknik *oversampling* tradisional (He, H., Bai, Y., Garcia, E., & Li, 2008). Eksperimen menggunakan berbagai *dataset* dengan tingkat ketidakseimbangan berbeda. Metrik evaluasi mencakup G-mean pada persamaan 2.58. Perbandingan statistik menggunakan Wilcoxon signed-rank test. Hasil menunjukkan peningkatan average *Precision*. ADASYN efektif untuk *dataset* dengan complex *decision* boundaries. Komputasi complexity adalah $O(n\log n)$ untuk n sampel. Implementasi praktis memerlukan tuning parameter β dan k. Integrasi dengan *Ensemble* methods meningkatkan *robust*nes.

2.3.5.3 *Undersampling* dengan *Tomek links*

Tomek (1976) mendefinisikan konsep *Tomek links* sebagai teknik *undersampling* untuk pembersihan *boundary*. Teknik ini mengidentifikasi dan menghapus pasangan sampel yang membingungkan di *decision boundary*. Dua instance (x_i, x_j) membentuk Tomek Link jika instance merupakan nearest neighbor satu sama lain. Syarat tambahannya adalah instance berasal dari kelas yang berbeda. Secara matematis, kondisi Tomek Link dinyatakan sebagaipersamaan 2.64. Fungsi NN(x) mengembalikan nearest neighbor dari instance x. Penghapusan *Tomek links* menyederhanakan *decision boundary* dengan menghilangkan *noise* (Tomek, 1976). Proses ini meningkatkan separability antara kelas mayoritas dan minoritas. Implementasi *Tomek links* efektif untuk *dataset* dengan overlapping Classes. Teknik ini biasanya dikombinasikan dengan metode *sampling* lainnya.

$$\text{TomekLink}(x_i, x_j) \iff \left(\text{NN}(x_i)=x_j\right) \wedge \left(\text{NN}(x_j)=x_i\right) \wedge \left(y_i \neq y_j\right) \quad (2.64)$$

Algoritma *Tomek links* bekerja melalui beberapa tahapan komputasi yang sistematis. Tahap pertama menghitung nearest neighbor untuk setiap instance dalam *dataset* (Tomek, 1976). Perhitungan distance menggunakan Euclidean metric pada persamaan 2.65. Tahap kedua mengidentifikasi semua pasangan instance yang saling menjadi nearest neighbor. Tahap ketiga memfilter pasangan yang berasal dari kelas berbeda. Tahap keempat menghapus instance dari kelas

mayoritas dalam setiap Tomek Link. Rasio penghapusan dihitung dengan persamaan 2.66. Proses ini mempertahankan structural integrity dari kelas minoritas. Kompleksitas algoritma adalah $O(n^2)$ untuk implementasi naif. Optimasi menggunakan k-d tree mengurangi kompleksitas menjadi $O(n\log n)$. Validasi hasil menggunakan visualisasi *decision boundary* sebelum dan sesudah.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{2.65}$$

$$\text{removal\_ratio} = \frac{n_{\text{tomek\_links}}}{n_{\text{total}}} \tag{2.66}$$

Evaluasi *Tomek links* menunjukkan efektivitas dalam meningkatkan kualitas *dataset* (Tomek, 1976). Eksperimen menggunakan multiple benchmark *dataset*s dengan berbagai karakteristik. Metrik purity *boundary* dihitung dengan persamaan 2.67. Integrated metric pada persamaan 2.68. *Statistical significance* diuji dengan paired t-test ($p < 0.05$). Implementasi praktis memerlukan pertimbangan Computational cost. Aplikasi successful dalam domain medical diagnosis dan text Classification.

$$P_{boundary} = 1 - \left(\frac{n_{\text{tomek\_links}}}{n_{boundary\_\text{samples}}}\right) \tag{2.67}$$

$$I_{\text{score}} = \alpha \times \text{G-mean} + \beta \times Precision_{\text{minority}} \tag{2.68}$$

2.3.6 *Retrieval-Augmented Generation* (RAG)

2.3.6.1 Arsitektur dan Komponen

Lewis et al. (2021) mendefinisikan *Retrieval-Augmented Generation* (RAG) sebagai *framework* terpadu untuk tugas-tugas NLP. *Framework* ini mengintegrasikan komponen retrieval dan generation dalam satu arsitektur. RAG mengatasi keterbatasan model generative murni dengan menyertakan informasi eksternal. Arsitektur RAG terdiri dari tiga komponen utama yang bekerja sinergis (Lewis et al., 2021). Komponen pertama adalah retriever yang bertugas mengambil dokumen relevan. Komponen kedua adalah Generator yang menghasilkan respons berdasarkan konteks. Komponen ketiga adalah *Knowledge Base* yang menyimpan informasi eksternal. Proses retrieval menggunakan dense vector Similarity dengan

persamaan 2.69. Generator mengikuti persamaan 2.70. Integrasi ini menghasilkan respons yang lebih akurat dan faktual.

$$\text{sim}(q, d) = \frac{q^\top d}{\|q\| \, \|d\|} \tag{2.69}$$

$$P(y \mid x) = \prod_i P(y_i \mid x, z, y_{1:i-1}) \tag{2.70}$$

Retriever component menggunakan dense passage retrieval (DPR) untuk efisiensi (Lewis et al., 2021). DPR mempelajari fungsi embedding: $E_Q(q)$ dan $E_D(d)$ ke ruang vektor yang sama. Similarity dihitung dengan dot product pada persamaan 2.71. *training Objective* memaksimalkan margin pada persamaan 2.72. *Knowledge Base* berisi koleksi dokumen pada persamaan 2.73. Top-k retrieval memilih dokumen dengan score tertinggi. Generator component biasanya menggunakan encoder-Decoder architecture. Contextualized representation dihitung dengan persamaan 2.74. Decoder menghasilkan token pada persamaan 2.75. End-to-end *training* memungkinkan joint *Optimization* kedua komponen.

$$\text{score}(q, d) = E_Q(q)^\top E_D(d) \tag{2.71}$$

$$\mathcal{L} = \max(0, \gamma - \text{score}(q, d^+) + \text{score}(q, d^-)) \tag{2.72}$$

$$Z = \{z_1, z_2, \ldots, z_n\} \tag{2.73}$$

$$h = Encoder([x; z]) \tag{2.74}$$

$$P(y_i) = \text{softmax}(W_o h_i + b_o) \tag{2.75}$$

Implementasi RAG menunjukkan peningkatan signifikan pada berbagai tugas NLP (Lewis et al., 2021). Evaluasi menggunakan *metrics*: Exact Match (EM) dan *F1-score* untuk QA tasks. Perhitungan EM pada persamaan 2.76. *F1-score* dihitung pada persamaan 2.77. Computational complexity pada $O(n)$ untuk retrieval dan $O(m^2)$ untuk generation. Optimasi menggunakan FAISS untuk efficient Similarity search. Aplikasi successful dalam domain question answering dan dialog systems.

$$\text{EM} = \begin{cases} 1, & \text{jika prediksi sama dengan ground truth} \\ 0, & \text{lainnya} \end{cases} \tag{2.76}$$

$$\text{F1} = 2 \times \frac{Precision \times recall}{Precision + recall} \tag{2.77}$$

2.3.6.2 RAG untuk Data *Distribution Alignment*

Guu et al. (2020) menGaplikasikan *framework* RAG untuk *distribution* alignment dalam pembelajaran mesin. Pendekatan ini mengatasi masalah ketidaksesuaian distribusi antara data *training* dan *testing*. *Distribution* alignment menjadi krusial untuk meningkatkan generalisasi model. RAG dimanfaatkan untuk mensintesis sampel yang match dengan distribusi data real. *Objective function* utama meminimalkan KL divergence: $\min_G D_{KL}(P_{\text{real}} \parallel P_{Synthetic})$. Proses ini mempertahankan karakteristik esensial dari data asli. *Framework* ini efektif untuk domain adaptation dan *Transfer learning*. Implementasi melibatkan joint *training* retriever dan Generator. Hasilnya adalah peningkatan konsistensi distribusional data sintetik. Aplikasi sukses dilaporkan pada domain medical imaging dan text Classification.

Distribution matching menggunakan adversarial *training* paradigm untuk alignment optimal (Guu et al., 2020). Generator G memproduksi sampel sintetik persamaan 2.78. Discriminator D membedakan real dan Synthetic samples persamaan 2.79. KL divergence term pada persamaan 2.80. Full *Objective function*: $\min_G \max_D (L_{\text{adv}} + \lambda D_{\text{KL}})$. *Hyperparameter* λ mengontrol *Trade-off* antara kualitas dan alignment. Retriever component menyediakan context untuk conditioning Generator. Conditioning vector c berasal dari retrieved documents pada persamaan 2.81. Proses *training* menggunakan alternating *Optimization* strategy. *Convergence* dicapai ketika $D_{\text{KL}} < \varepsilon$ threshold. Monitoring menggunakan *distribution* distance *metrics* secara berkala.

$$\tilde{x} = G(z, c) \tag{2.78}$$

$$L_{\text{adv}} = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(\tilde{x}))] \tag{2.79}$$

$$D_{\text{KL}} = \mathbb{E}\left[P_{\text{real}} \log\left(\frac{P_{\text{real}}}{P_{Synthetic}}\right)\right] \tag{2.80}$$

$$\text{c} = \text{R(q)} \tag{2.81}$$

Knowledge retrieval memainkan peran vital dalam *distribution* alignment process (Gao et al., 2024). Query persamaan 2.82 untuk input instance x. Document

retrieval pada persamaan 2.83. Context aggregation pada persamaan 2.84. Similarity *function* pada persamaan 2.85. Retriever *training* menggunakan contrastive *learning Objective*. *Positive* pairs: (q, z^+) dari relevant documents. Negative pairs: (q, z^-) dari irrelevant documents. *Loss function* pada persamaan 2.86. Fine-tuning retriever meningkatkan quality retrieved contexts. Optimal retrieval meningkatkan alignment *accuracy* secara signifikan.

$$q = f_{\text{enc}}(x) \tag{2.82}$$

$$Z = \text{top-}k\left(\underset{z}{\text{argmax}}\, \text{sim}(q,z)\right) \tag{2.83}$$

$$c = g_{\text{agg}}(z_1, \dots, z_k) \tag{2.84}$$

$$\text{sim}(q, z) = \frac{\exp(q^\top z)}{\sum_j \exp(q^\top z_j)} \tag{2.85}$$

$$L_{\text{ret}} = -\log\left[\frac{\exp(\text{sim}(q,z^+))}{\sum_j \exp(\text{sim}(q,z_j))}\right] \tag{2.86}$$

Generator architecture mengintegrasikan retrieved contexts untuk Synthetic sample generation (Gao et al., 2024). Conditional generation pada persamaan 2.87. Context integration pada persamaan 2.88. Output *distribution* pada persamaan 2.89. *training Objective* pada persamaan 2.90. *Regularization* term pada persamaan 2.91. Total *loss* pada persamaan 2.92. *Hyperparameter*s α dan β di-tune via cross-*Validation*. *Sampling Diversity* dikontrol via *Temperature parameter* τ. Beam search meningkatkan coherence generated samples. Quality *Assessment* menggunakan Automated *metrics* dan human *evaluation*.

$$P(\tilde{x} \mid c) = \prod_t P(\tilde{x}_t \mid \tilde{x}_{<t}, c) \tag{2.87}$$

$$h_t = \textit{Transformer}(\tilde{x}_{<t}, c) \tag{2.88}$$

$$P(\tilde{x}_t) = \text{softmax}(W_o h_t + b_o) \tag{2.89}$$

$$L_{\text{gen}} = -\mathbb{E}[\log P(\tilde{x} \mid c)] \tag{2.90}$$

$$L_{\text{reg}} = \| \theta_G \|_2^2 \tag{2.91}$$

$$L_{\text{total}} = L_{\text{gen}} + \alpha L_{\text{adv}} + \beta L_{\text{reg}} \tag{2.92}$$

2.3.7 Evaluasi Distribusi dan Kinerja Model

Evaluasi kinerja merupakan tahap krusial dalam pengembangan model klasifikasi. Proses memastikan keakuratan dan keandalan model dalam melakukan prediksi. Para Penelitian menggunakan berbagai metrik evaluasi umum dan terstandarisasi. Metrik akurasi mengukur persentase prediksi benar dari keseluruhan data uji. Metrik presisi menghitung proporsi prediksi positif benar dari semua hasil diprediksi positif (Dash et al., 2023). Sementara itu, metrik *Recall* menunjukkan kemampuan model menemukan semua sampel positif ada. *F1-score* menjadi rata-rata harmonik menggabungkan presisi dan *Recall*. Metrik memberikan keseimbangan antara *False Positive* dan *False Negative* (Rosline Mary & Kavitha, 2023). Setiap metrik menawarkan perspektif berbeda tentang performa model. Kesimpulan akhir berasal dari analisis komprehensif terhadap semua metrik.

Akurasi tidak menjadi tolak ukur memadai untuk data tidak seimbang. Model dapat mencapai skor akurasi tinggi hanya dengan memprediksi kelas mayoritas. Evaluasi performa memerlukan tiga metrik kunci saling melengkapi. *True Positive* (TP) mencatat keberhasilan model dalam mengidentifikasi kelas minoritas dengan benar (Sokolova & Lapalme, 2009). *False Positive* (FP) menunjukkan kesalahan model dalam menandai kelas mayoritas sebagai minoritas. *False Negative* (FN) mengukur kegagalan model dalam mendeteksi contoh kelas minoritas aktual. Nilai Presisi menghitung rasio prediksi minoritas benar dari seluruh prediksi positif. Nilai *Recall* menunjukkan proporsi kasus minoritas berhasil diidentifikasi oleh model. *F1-score* merupakan rata-rata harmonik menggabungkan Presisi dan *Recall* (Ramos et al., 2023). Teknik ADASYN dan *Tomek links* dapat meningkatkan ketiga metrik secara signifikan.

Metrik evaluasi memberikan gambaran komprehensif tentang performa suatu model. Berbagai metrik mengukur kemampuan prediksi model untuk kelas positif dan negatif. Setiap metrik menunjukkan aspek berbeda dari kinerja model secara keseluruhan. Analisis mendalam mengungkap kekuatan dan kelemahan model dalam klasifikasi. Metrik utama membantu menghindari kesalahan seperti *False Positive* dan *False Negative*. Penggunaan beberapa metrik memastikan

penilaian adil dan tidak bias. Penelitian dapat mengambil keputusan berdasarkan pemahaman utuh. Evaluasi baik meningkatkan kepercayaan terhadap hasil diproduksi model. Pendekatan komprehensif menjadi standar dalam menilai model klasifikasi (Yu et al., 2024). Hasil evaluasi menjadi dasar untuk perbaikan dan pengembangan model lebih lanjut.

Akurasi merupakan metrik evaluasi paling dasar dan intuitif. Metrik mengukur persentase prediksi benar dari keseluruhan data uji. Akurasi memberikan gambaran umum tentang performa model secara keseluruhan. Nilainya menunjukkan sejauh mana model dapat membuat prediksi tepat. Perhitungan akurasi mencakup prediksi benar untuk kedua kelas, baik positif maupun negatif. Metrik menjadi indikator awal mudah dipahami oleh berbagai pihak. Namun, akurasi memiliki keterbatasan tertentu pada situasi data tidak seimbang. Para Penelitian harus melengkapi akurasi dengan metrik lain lebih spesifik. Akurasi tetap menjadi patokan penting dalam laporan kinerja model. Penjelasan mengenai persamaan 2.93 perhitungannya terdapat pada bagian metodologi.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.93)$$

Evaluasi performa model klasifikasi menggunakan empat metrik fundamental saling melengkapi. *True Positive* (TP) merekam jumlah prediksi benar untuk kelas positif berhasil diidentifikasi model (Wojarnik, 2021). *True Negative* (TN) mengukur keberhasilan model dalam mengenali contoh kelas negatif dengan tepat. *False Positive* (FP) mencatat kesalahan model ketika memprediksi contoh negatif sebagai positif. *False Negative* (FN) menghitung jumlah kasus dimana model gagal mendeteksi contoh positif sebenarnya. Keempat komponen membentuk dasar perhitungan berbagai metrik evaluasi lebih kompleks. Rasio TP terhadap total prediksi positif menghasilkan nilai presisi mencerminkan ketepatan prediksi. Pembagian TP dengan jumlah aktual positif memberikan nilai *Recall* mengukur kelengkapan deteksi. Kombinasi presisi dan *Recall* melalui harmonic mean menghasilkan skor F1 sebagai indikator keseimbangan performa (Góra & Skowron, 2025). Analisis menyeluruh terhadap matriks konfusi memungkinkan

evaluasi komprehensif terhadap kelebihan dan kekurangan model.

Akurasi merupakan metrik evaluasi sangat efektif untuk *dataset* seimbang. Metrik memberikan gambaran andal tentang performa model secara keseluruhan. Namun, akurasi dapat menyesatkan pada kasus ketidakseimbangan data tinggi. Situasi sering terjadi ketika satu kelas jauh lebih dominan daripada kelas lainnya. Model mungkin mencapai akurasi tinggi hanya dengan memprediksi kelas mayoritas (Çalişir & Doğantekin, 2011). Akurasi tinggi tidak mencerminkan kemampuan model sesungguhnya. Kelas minoritas seringkali justru lebih kritis untuk dideteksi dengan benar. Para Penelitian harus menggunakan metrik tambahan untuk evaluasi lebih adil. Metrik seperti presisi, *Recall*, dan *F1-score* memberikan wawasan lebih mendalam. Pendekatan evaluasi komprehensif menjamin penilaian kinerja model lebih akurat.

Presisi mengukur proporsi prediksi positif benar dari seluruh prediksi positif model. Metrik menjadi kriteria evaluasi sangat penting dalam banyak aplikasi. Beberapa bidang memiliki toleransi kesalahan sangat rendah untuk *False Positive*. Diagnosis medis merupakan contoh dimana kesalahan prediksi positif berakibat fatal. Deteksi penipuan memerlukan tingkat presisi sangat tinggi. Kesalahan klasifikasi dapat menimbulkan kerugian material dan non-material besar. Presisi menjadi indikator kunci untuk mengukur keandalan prediksi positif. Nilai presisi tinggi mencerminkan ketepatan model dalam memprediksi kelas positif. Para Penelitian harus memperhatikan metrik ketika *False Positive* tidak dapat diterima. Persamaan 2.94 untuk menghitung presisi tersedia dalam bagian metodologi penelitian.

$$\text{Presisi} = \frac{TP}{TP+FP} \tag{2.94}$$

Nilai presisi tinggi menunjukkan performa model jarang melakukan kesalahan prediksi positif. Model hanya memprediksi suatu *instance* sebagai positif ketika memiliki keyakinan tinggi. Namun, pencapaian bisa berasal dari kecenderungan model terlalu selektif. Situasi sering menghasilkan nilai *Recall* rendah secara bersamaan. *Recall* rendah mengindikasikan bahwa model gagal

mendeteksi banyak *instance* positif sebenarnya (Ainapure et al., 2023). Model mungkin melewatkan banyak kasus positif seharusnya teridentifikasi. Kombinasi presisi tinggi dan *Recall* rendah menciptakan gambaran performa tidak lengkap. Evaluasi model memerlukan pertimbangan kedua metrik secara bersamaan. Penelitian harus menganalisis *Trade-off* antara presisi dan *Recall*. Keseimbangan antara kedua metrik menentukan efektivitas model secara keseluruhan.

Recall mengukur kemampuan model dalam mengidentifikasi seluruh *instance* positif ada dalam *dataset* (Rahman et al., 2019). Metrik menjadi sangat kritis dalam berbagai aplikasi penting. Bidang diagnosis medis memprioritaskan *Recall* tinggi untuk mendeteksi semua kasus penyakit. Sistem keamanan membutuhkan *Recall* optimal untuk mencegah ancaman terlewat. *Recall* menghitung proporsi *instance* positif berhasil diprediksi dengan benar (Li et al., 2024). Nilai *Recall* rendah menunjukkan banyaknya *False Negative* dalam prediksi model. Kondisi dapat berakibat sangat serius dalam skenario dunia nyata. Model dengan *Recall* tinggi memastikan deteksi komprehensif terhadap kelas positif. Para ahli menggunakan rumus tertentu untuk menghitung nilai *Recall* secara akurat. Detail persamaan 2.95 merupakan *Recall* tersedia pada bagian metodologi dalam laporan penelitian.

$$Recall = \frac{TP}{TP+FN} \tag{2.95}$$

Nilai *Recall* tinggi menunjukkan kemampuan model sangat baik dalam mendeteksi *instance* positif (Azmi et al., 2025). Model berhasil mengidentifikasi hampir semua kasus positif ada dalam *dataset.* Pencapaian sangat krusial untuk aplikasi tidak boleh melewatkan kasus positif. Namun, pencapaian *Recall* tinggi sering datang dengan konsekuensi tertentu. Model mungkin menjadi terlalu agresif dalam memprediksi kelas positif. Kondisi menyebabkan peningkatan jumlah *False Positive* signifikan. Akibatnya, nilai presisi bisa menjadi rendah secara bersamaan. Banyak prediksi positif ternyata merupakan kesalahan klasifikasi. Situasi menciptakan *Trade-off* perlu dipertimbangkan secara hati-hati. Keseimbangan antara *Recall* dan presisi menentukan kualitas model secara keseluruhan.

F1-score merupakan rata-rata harmonik menggabungkan presisi dan *Recall*. Metrik memberikan gambaran seimbang tentang performa model klasifikasi (Bhowmik et al., 2022). *F1-score* menjadi alat evaluasi sangat berguna untuk data tidak seimbang. Kombinasi mempertimbangkan kedua aspek prediksi secara simultan. Metrik membantu menghindari kesalahan *False Positive* dan *False Negative*. Nilai *F1-score* mencerminkan keseimbangan antara ketepatan dan kelengkapan deteksi. Para Penelitian sering mengutamakan metrik dalam evaluasi model. *F1-score* menjadi indikator lebih komprehensif daripada akurasi. Perhitungan metrik menggunakan rumus matematis spesifik. Detail persamaan 2.96 merupakan *F1-score* pada bagian lampiran penelitian.

$$F1 = \frac{2 \cdot \text{Presisi} \cdot \textit{Recall}}{\text{Presisi} + \textit{Recall}} \tag{2.96}$$

F1-score memberikan nilai optimal ketika model menyeimbangkan presisi dan *Recall* dengan baik. Metrik mencapai puncaknya ketika kedua komponen memiliki nilai seimbang. Perbedaan besar antara presisi dan *Recall* menyebabkan penurunan nilai *F1-score* secara signifikan. Rata-rata harmonik bersifat lebih sensitif terhadap ketidakseimbangan dibandingkan rata-rata aritmatika. Sensitivitas menjadikan *F1-score* sebagai indikator baik untuk mendeteksi bias model (Chatterjee et al., 2021). Nilai F1 rendah mengindikasikan bahwa model perlu penyesuaian pada salah satu aspek. Model harus dioptimalkan untuk mencapai keseimbangan antara kedua metrik. Penelitian dapat menggunakan *F1-score* sebagai panduan dalam menyempurnakan model. Metrik secara efektif mencegah pengabaian terhadap salah satu jenis error. Keseimbangan sangat penting untuk kinerja model baik secara keseluruhan.