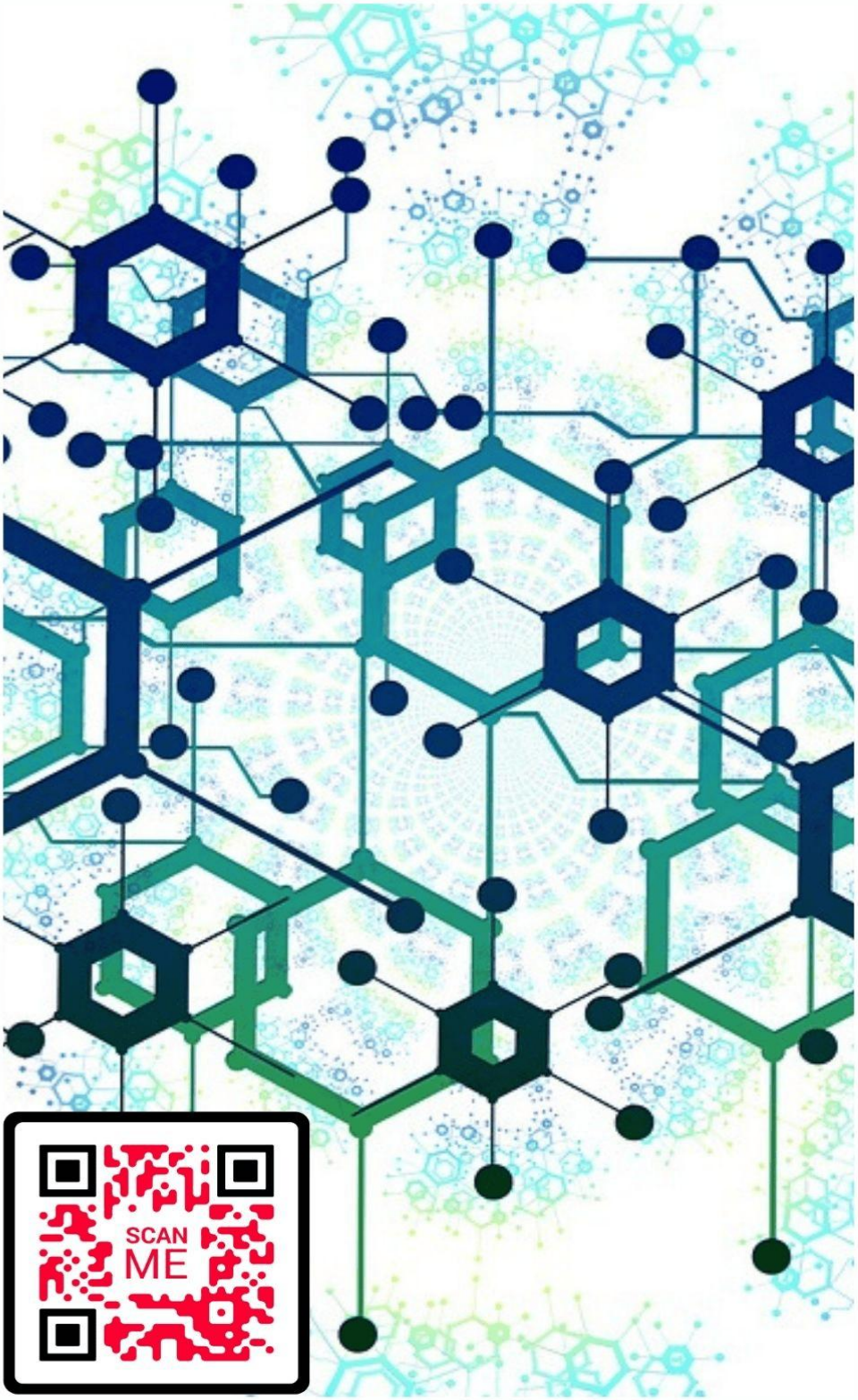
Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

ISSN (ONLINE) 2598-9936



INDONESIAN JOURNAL OF INNOVATION STUDIES

PUBLISHED BY

UNIVERSITAS MUHAMMADIYAH SIDOARJO

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Table Of Contents

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

EDITORIAL TEAM

Indonesian Journal of Innovation Studies

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

Indonesian Journal of Innovation Studies

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Generative AI Label Reliability and Text Classification in Programming Assessment

Reliabilitas Label Generative AI dan Klasifikasi Teks pada Asesmen Algoritma Pemrograman

Raissa Araminta Khairunnisa, raissa.araminta.2205356@students.um.ac.id, (1)
Program Studi Teknik Informatika, Universitas Negeri Malang, Indonesia

Utomo Pujianto, utomo.pujianto.ft@um.ac.id, ()
Program Studi Teknik Informatika, Universitas Negeri Malang, Indonesia

(1) Corresponding author

Abstract

General Background: The integration of Generative AI in educational assessment enables rapid construction of large-scale question banks, particularly in programming education, yet raises concerns regarding content validity. **Specific Background**: In algorithm and programming domains, Generative AI models frequently assign Higher Order Thinking Skills and Lower Order Thinking Skills labels automatically, creating potential discrepancies with Bloom's Taxonomy classifications. **Knowledge Gap**: Empirical evidence validating the reliability of AI-generated cognitive labels and comparing statistical and transformer-based classification methods on small, domain-specific Indonesian datasets remains limited. **Aims**: This study aims to audit the reliability of cognitive labels generated by the Gemini model through expert validation and to compare TF-IDF–SVM and IndoBERT–SVM classifiers under class-imbalanced conditions. **Results**: Expert validation revealed substantial mislabeling, with a claimed balanced dataset becoming skewed toward LOTS. Classification experiments using five-fold cross-validation showed that TF-IDF–SVM achieved a slightly higher macro F1-score than IndoBERT–SVM. **Novelty**: The study demonstrates that simple lexical representations with stemming can outperform transformer-based embeddings when data are limited and domain-specific. **Implications**: These findings emphasize the necessity of human validation in AI-generated assessments and support the use of lightweight statistical text classification for automated cognitive level evaluation in constrained educational contexts.

Highlights

- Generative AI cognitive labels showed substantial inconsistency after expert validation
- Lexical feature representation yielded higher macro-level classification balance
- Human-in-the-loop validation remained essential for programming assessment datasets

Keywords

HOTS; LOTS; Generative AI; Text Classification; TF-IDF

Published date: 2026-01-04

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

I. Pendahuluan

Integrasi teknologi kecerdasan buatan (Artificial Intelligence) dalam ekosistem pendidikan telah berkembang pesat, khususnya dengan kehadiran Generative AI seperti Google Gemini. Teknologi ini menawarkan efisiensi preseden baru dalam penyusunan instrumen evaluasi, di mana pendidik kini dapat menyusun ratusan butir soal latihan dalam waktu singkat, mengatasi kendala biaya dan waktu yang selama ini menghambat pengembangan item tes secara manual [1], [2]. Fenomena ini menjawab kebutuhan mendesak akan ketersediaan bank soal yang masif dan bervariasi, yang sangat krusial untuk mendukung kelancaran sistem pembelajaran adaptif serta mengatasi menipisnya stok soal akibat penggunaan asesmen yang intensif [3], [4].

Namun, di balik efisiensi tersebut, terdapat tantangan validitas konten yang serius, khususnya dalam akurasi klasifikasi tingkat kognitif yang didasarkan pada revisi Taksonomi Bloom. Kerangka ini mengkategorikan soal menjadi keterampilan berpikir tingkat rendah (LOTS) yang meliputi dimensi mengingat (C1) hingga menerapkan (C3), dan keterampilan berpikir tingkat tinggi (HOTS) yang melibatkan ranah menganalisis (C4), mengevaluasi (C5), dan mencipta (C6) [5]. Permasalahan mendasar muncul ketika AI mengalami fenomena "halusinasi", yaitu kecenderungan inheren model generatif untuk memproduksi konten yang secara linguistik fasih, namun secara faktual salah atau tidak berdasar pada logika dunia nyata [6], [7]. Dalam konteks pembuatan item tes, bias ini sering kali bermanifestasi ketika AI menghasilkan soal yang diklaim sesuai dengan level kognitif tinggi, namun pada kenyataannya gagal menyelaraskan kedalaman kognitif dengan target Taksonomi Bloom yang diminta [8], [9]. Inkonsistensi ini berisiko menghasilkan instrumen penilaian yang cacat validitas konstruksinya, sehingga mutlak memerlukan validasi manusia untuk memastikan akurasi pemetaan kemampuan berpikir kritis siswa [1], [9].

Dalam upaya mendeteksi kesalahan pelabelan tersebut secara otomatis, studi-studi terdahulu umumnya mengadopsi pendekatan klasifikasi teks berbasis statistik leksikal, seperti menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) yang dikenal karena efisiensinya. Prinsip metode ini adalah menetapkan bobot numerik kata untuk merefleksikan signifikansi kata dalam sebuah dokumen dibandingkan dengan seluruh koleksi korpus [10]. Meskipun cepat dan intuitif, literatur menunjukkan bahwa representasi berbasis bag-of-words seperti TF-IDF memiliki batasan teoretis yang signifikan karena sifatnya yang sparse dan kegagalannya dalam menangkap sinonim serta polisemi kata, yang pada akhirnya mengabaikan urutan dan makna semantik antar kata [11]. Hal ini memunculkan pertanyaan penelitian apakah metode statistik sederhana ini cukup mumpuni untuk menangani ambiguitas klasifikasi yang membutuhkan pemahaman konteks semantik yang mendalam.

Sebagai alternatif pendekatan statistik, perkembangan Deep Learning menawarkan arsitektur Transformer yang mampu memahami bahasa secara mendalam. Model yang paling relevan untuk konteks Indonesia adalah IndoBERT, sebuah varian BERT yang dilatih khusus menggunakan korpus masif Bahasa Indonesia untuk memastikan keragaman gaya bahasa [12]. Kemampuan IndoBERT yang mengadopsi mekanisme Bidirectional Encoder Representations diklaim lebih unggul dalam menangkap nuansa semantik yang sering kali tersirat dalam soal HOTS dan LOTS [13]. Namun, meskipun model ini menawarkan kinerja state-of-the-art pada berbagai tugas pemahaman bahasa, efektivitas praktisnya pada dataset berskala kecil masih perlu diuji secara empiris dibandingkan metode baseline yang lebih sederhana.

Kesenjangan penelitian utama terletak pada asumsi yang sering diterima bahwa dataset yang dihasilkan AI akurat tanpa melalui validasi mendalam, padahal sistem penilaian otomatis sering kali gagal menginterpretasikan nuansa konteks yang kompleks [14]. Secara spesifik, belum banyak penelitian yang mengevaluasi akurasi label AI pada domain "Algoritma Pemrograman" dengan menggunakan penilaian pakar manusia sebagai standar kebenaran utama (Ground Truth) guna mengoreksi potensi kesalahan pelabelan tersebut. Selain itu, efektivitas model pre-trained berbasis transformer pada data berskala kecil dan spesifik domain masih diperdebatkan secara empiris [15]. Berdasarkan permasalahan yang diuraikan, penelitian ini memiliki dua tujuan utama. Tujuan pertama adalah mengevaluasi reliabilitas label otomatis yang dihasilkan oleh model Gemini pada 500 soal algoritma pemrograman melalui mekanisme validasi pakar (human-in-the-loop). Tujuan kedua adalah melakukan analisis komparatif antara model klasifikasi hybrid IndoBERT-SVM melawan baseline TF-IDF guna mengidentifikasi pendekatan yang menawarkan efisiensi sekaligus efektivitas tertinggi dalam penanganan dataset yang bias.

II. Metode

Penelitian ini menerapkan pendekatan eksperimental kuantitatif untuk mengevaluasi reliabilitas data sintesis AL dan menguji efektivitas model klasifikasi dalam menangani ketidakseimbangan kelas. Struktur system yang dikembangkan mencakup empat tahapan pokok yang saling terintegrasi, yaitu pengumpulan data sintesis, validasi pakar, pra-pemrosesan teks, dan komprasi model klasifikasi. Proses kerja system secara menyeluruh dapat diamati pada Gambar 1, yang menggambarkan perjalanan data dari output mentah Gemini hingga menjadi prediksi kelas oleh model.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

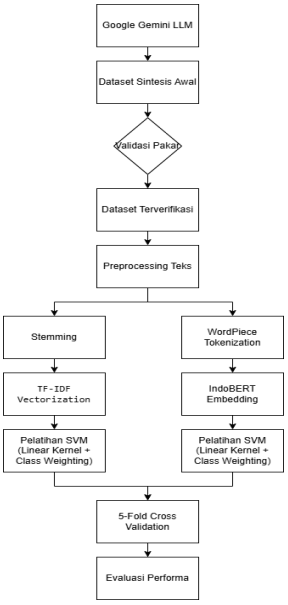


Figure 1. *Alur Kerja Sistem Penelitian*

a. Pengumpulan Data Sintesis dan Validasi Pakar

Tahap awal penelitian difokuskan pada pengumpulan dataset menggunakan Large Language Model (LLM) Google Gemini. Instruksi spesifik (prompt) diberikan kepada model untuk menyusun 500 butir soal pilihan ganda pada domain "Algoritma dan Pemrograman", sebuah pendekatan yang terbukti mampu menghasilkan item penilaian berkualitas tinggi yang selaras dengan kerangka pendidikan seperti Taksonomi Bloom [8]. Namun, untuk menjamin kualitas data sebelum digunakan dalam pelatihan model klasifikasi, seluruh data mentah melalui tahapan validasi manual oleh pakar materi (human expert), mengingat studi terdahulu menegaskan bahwa output AI sering kali memerlukan intervensi manusia untuk mengoreksi inakurasi faktual atau bias algoritma [1], [14].

Validasi ini dilaksanakan oleh dua orang dosen ahli dari Program Studi Teknik Informatika, yang memiliki spesialisasi dan pengalaman mengajar yang relevan pada mata kuliah Algoritma dan Pemrograman. Dalam studi sejenis, konsistensi antar pakar biasanya diukur dengan koefisien statistic seperti Kohen's Kappa (Inter-rater reliability/IRR). Namun, dalam penelitian ini, perhitungan IRR seperti Kohen's Kappa tidak dapat diterapkan. Hal ini disebabkan oleh desain pembagian tugas validasi, di mana setiap pakar menilai set soal yang berbeda secara keseluruhan dan tidak saling tumpang tindih, demi efisiensi waktu dan sumber daya. Sebagai alternatif mitigasi, konsistensi internal dijaga melalui penggunaan pedoman validasi yang terstandarisasi dan eksplisit, yaitu kesesuaian narasi soal dengan Kata Kerja Operasional (KKO) Taknosomi Bloom revisi. Dengan demikian, keputusan label akhir (Ground Truth) diserahkan sepenuhnya kepada penilaian pakar yang ditugaskan untuk set soal tersebut, dengan asumsi bahwa kedua pakar memiliki pemahaman yang setara terhadap pedoman yang diberikan. Proses validasi dilakukan dengan meninjau kesesuaian narasi soal terhadap KKO Taksonomi Bloom untuk menetapkan Ground Truth, langkah yang krusial karena model generatif terkadang gagal menyelaraskan kedalaman kognitif soal dengan tingkat kesulitan yang ditargetkan [8]. Setelah melalui proses verifikasi dan pembersihan data, diperoleh dataset final yang valid sebanyak 500 butir soal dengan komposisi 178 soal kategori HOTS dan 322 soal kategori LOTS. Dataset inilah yang selanjutnya diproses ke tahap pra-pemrosesan teks.

b. Pra-pemrosesan Teks (Preprocessing)

Data yang telah divalidasi kemudian melalui fase pra-pemrosesan untuk membersihkan serta menstandarkan teks agar mampu direpresentasikan oleh mesin. Tahapan ini dimulai dengan case folding untuk menyeragamkan seluruh huruf menjadi huruf kecil, dilanjutkan dengan cleaning untuk menghapus karakter non-alfanumerik seperti tanda baca dan angka yang tidak relevan dengan konteks pemrograman. Selanjutnya, dilakukan stopword removal menggunakan pustaka Sastrawi untuk mengeliminasi kata-kata umum (seperti "yang", "dan", "dari") yang memiliki frekuensi kemunculan tinggi namun memberikan kontribusi yang sangat minim terhadap makna semantik atau isi dokumen, sehingga penghapusannya dianggap penting untuk meningkatkan efisiensi representasi teks [16], [17].

Perbedaan perlakuan diterapkan secara spesifik pada tahap representasi kata (tokenisasi) sesuai kebutuhan arsitektur model. Untuk skenario model TF-IDF, dilakukan proses stemming berbasis algoritma Nazief & Adriani guna mereduksi kata berimbuhan menjadi kata dasar (misalnya "memprogram" menjadi "program"). Hal ini bertujuan untuk mengurangi dimensi fitur (sparsity) pada matriks TF-IDF, mengingat representasi berbasis bag-of-words cenderung menghasilkan vektor yang sangat jarang dan gagal menangkap polisemi antar kata [11]. Sebaliknya, untuk skenario model IndoBERT, proses stemming ditiadakan karena dapat menghilangkan informasi gramatikal yang penting. Sebagai gantinya, digunakan WordPiece Tokenizer yang memecah kata kompleks menjadi unit sub-kata yang lebih kecil (misalnya "pemrograman" menjadi "pem", "rog", "raman") menggunakan strategi maximum matching [18]. Teknik representasi sub-kata ini memungkinkan model BERT untuk mengenali istilah teknis yang jarang muncul atau tidak ada dalam kamus (out-of-vocabulary) secara lebih efektif

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

dibandingkan tokenisasi tingkat kata tradisional, sekaligus menjaga keutuhan konteks kalimat [19]. Ilustrasi perbandingan hasil pra-pemrosesan antara kedua skenario tersebut dapat dilihat pada Tabel 1.

Tahapan	Contoh Teks
Teks Asli	Manakah dari pernyataan berikut yang paling tepat menggambarkan perbedaan utama antara antarmuka baris perintah (CLI) dan antarmuka pengguna grafis (GUI)?
TF-IDF (Stemming)	"mana nyata ikut paling tepat gambar beda utama antarmuka baris perintah cli antarmuka guna grafis gui"
IndoBERT (Tokenizer)	[CLS] manakah pernyataan berikut paling tepat menggambarkan perbedaan utama antarmuka baris perintah cli antarmuka pengguna grafis gui [SEP]

Table 1. *Ilustrasi Perbandingan Hasil Pra-Pemrosesan Antara Kedua Skenario*

c. Model Klasifikasi

Penelitian ini membandingkan dua skenario pemodelan untuk menguji hipotesis efektivitas representasi fitur terhadap akurasi klasifikasi. Meskipun kedua skenario menggunakan algoritma klasifikasi akhir yang sama, yaitu Support Vector Machine (SVM) dengan kernel yang digunakan adalah Linear, perbedaan fundamental terletak pada paradigma transformasi teks menjadi representasi numerik (vektor).Skenario pertama menggunakan pendekatan statistik tradisional Term Frequency-Inverse Document Frequency (TF-IDF). Dalam skenario ini, sistem membangun matriks kosa kata (vocabulary) berdasarkan frekuensi kemunculan kata. Prinsip kerjanya adalah memberikan bobot tinggi pada kata yang sering muncul dalam satu soal spesifik (Term Frequency) namun jarang muncul di dokumen lain (Inverse Document Frequency), sehingga kata tersebut diklasifikasikan sebagai kunci unik pembeda kelas yang signifikan secara statistik [10]. Secara matematis, pembobotan ini dilakukan melalui tiga tahapan perhitungan. Pertama, nilai Term Frequency (TF) dihitung menggunakan Persamaan (1).

$$TF = \frac{x}{y} \quad (1)$$

Di mana x adalah frekuensi kemunculan kata dalam dokumen, dan y adalah total jumlah kata di dalam dokumen tersebut. Selanjutnya, nilai Inverse Document Frequency (IDF) dihitung menggunakan Persamaan (2).

$$IDF = \log \binom{i}{j} \quad (2)$$

Di mana i adalah jumlah keseluruhan dokumen dalam korpus dan j adalah jumlah dokumen di mana kata tersebut muncul setidaknya satu kali. Akhirnya, bobot representasi fitur TF-IDF diperoleh dengan mengalikan kedua komponen tersebut sebagaimana ditunjukkan pada Persamaan (3).

$$TF - IDF = TF \times IDF \quad (3)$$

Kelemahan mendasar dari pendekatan ini adalah sifatnya yang bag-of-words, di mana urutan kata diabaikan dan model tidak dapat mengenali istilah yang tidak ada dalam kamus latih (Out-of-Vocabulary). Vektor yang dihasilkan bersifat sparse atau jarang, di mana sebagian besar elemen bernilai nol karena kegagalan metode ini dalam menangkap sinonim atau polisemi antar kata [11]. Vektor bobot inilah yang kemudian menjadi input bagi SVM untuk mencari hyperplane pemisah.

Skenario kedua adalah model usulan Hybrid IndoBERT-SVM. Skenario ini menggantikan pendekatan statistik dengan pendekatan semantik menggunakan model pre-trained indobert-base-p1 sebagai ekstraktor fitur statis (frozen feature extractor) [12]. Berbeda dengan TF-IDF yang memecah kalimat menjadi kata terpisah, IndoBERT memproses kalimat sebagai satu kesatuan urutan (sequence) menggunakan mekanisme Self-Attention. Mekanisme ini memungkinkan model untuk mempertimbangkan keterkaitan antar kata, misalnya bagaimana kata sebelumnya mengubah makna kata berikutnya. Setiap soal diproses melewati 12 lapisan encoder, dan output yang diambil adalah vektor embedding dari token spesial [CLS] pada lapisan terakhir. Token [CLS] ini berfungsi sebagai representasi agregat yang merangkum makna semantik keseluruhan kalimat, berbeda dengan token kata biasa [13]. Vektor ini memiliki dimensi tetap sebesar 768 dan bersifat padat (dense vector), yang berisi informasi kontekstual kaya untuk diproses lebih lanjut oleh SVM.

Peran SVM pada kedua skenario tersebut adalah sebagai classifier yang bertugas menentukan batas keputusan (decision boundary) optimal. Dalam penelitian ini, fungsi kernel yang diterapkan adalah Kernel Linear. Pemilihan kernel ini didasarkan pada karakteristik data teks berdimensi tinggi (seperti TF-IDF dan Embedding IndoBERT) yang cenderung sudah terpisah secara linier, sehingga penggunaan model yang lebih sederhana sering kali memberikan efisiensi komputasi yang lebih baik tanpa mengorbankan performa dibandingkan model yang lebih kompleks [20]. Secara matematis, fungsi Kernel

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Linear didefinisikan sebagai perkalian titik (dot product) sederhana antara dua vektor fitur, sebagaimana ditunjukkan pada Persamaan (4).

$$K(x_i, x_j) = X_i^T X_j \quad (4)$$

Di mana X_i dan X_j merepresentasikan pasangan vektor dokumen dalam ruang fitur. Pada skenario TF-IDF, SVM memisahkan data berdasarkan pencocokan kata kunci leksikal, sedangkan pada skenario IndoBERT, SVM memisahkan data berdasarkan klaster makna di ruang vektor. Penggunaan kernel linear dipilih secara konsisten di kedua skenario untuk memastikan bahwa perbedaan performa yang dihasilkan murni berasal dari kualitas fitur, bukan kompleksitas kernel. Selain itu, guna menangani tantangan ketidakseimbangan data yang signifikan, penelitian ini menerapkan teknik Class Weighting (atau dikenal sebagai Cost-Sensitive Learning). Teknik ini bekerja dengan memodifikasi fungsi objektif SVM untuk memberikan bobot penalti yang lebih besar jika model melakukan kesalahan prediksi pada kelas minoritas (HOTS) dibandingkan kelas mayoritas (LOTS). Pendekatan ini memungkinkan hyperplane keputusan bergeser ke arah yang menguntungkan kelas minoritas, sehingga model menjadi lebih sensitif terhadap ciri-ciri soal HOTS tanpa perlu memanipulasi jumlah data fisik melalui resampling [21].

d. Evaluasi Model

Proses pengujian keandalan model dilakukan menggunakan protokol K-Fold Cross Validation dengan nilai $k = 5$. Pemilihan metode ini didasarkan pada keterbatasan jumlah dataset (500 butir soal) yang rentan terhadap bias varians jika hanya menggunakan metode pembagian data latih-uji statis (hold-out). Dengan K-Fold, dataset dibagi menjadi lima bagian (fold) proporsional, di mana setiap bagian akan mendapatkan giliran sebagai data uji sementara empat bagian lainnya menjadi data latih. Mekanisme rotasi ini memastikan bahwa setiap butir soal pernah dievaluasi sebagai data uji, sehingga metric performa yang dihasilkan mencerminkan kemampuan generalisasi model yang sebenarnya dan bukan sekedar kebetulan akibat pembagian data yang menguntungkan.

Perhitungan dasar untuk mengukur kinerja model menggunakan matriks kebigungan (Confusion Matrix), yang berfungsi memetakan hasil prediksi model dengan mengelompokkannya ke dalam empat kategori, yaitu: True positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Penelitian ini tidak menggunakan akurasi global sebagai tolak ukur utama karena potensi accuracy paradox pada data tidak seimbang, di mana model dapat mencapai skor tinggi hanya dengan memprediksi kelas mayoritas (LOTS) namun gagal mengenali kelas minoritas (HOTS) yang justru menjadi fokus utama deteksi. Oleh karena itu, evaluasi difokuskan pada tiga metrik spesifik. Pertama adalah Precision (Persamaan 5) untuk mengukur ketepatan prediksi, dan kedua adalah Recall (Persamaan 6) untuk mengukur sensitivitas model terhadap kelas tertentu [22].

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \quad (5, 6)$$

Metrik ketiga yang menjadi acuan penentu dalam penelitian ini adalah F1-Score. Nilai F1-Score didapatkan dari rata-rata harmonik antara Precision dan Recall, dan ini menunjukkan keseimbangan kinerja model dari aspek ketepatan serta sensitivitas. Perhitungan nilai ini dilakukan menggunakan Persamaan (7) [22].

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Mengingat adanya ketidakseimbangan jumlah data antara kelas HOTS dan LOTS, nilai akhir performa model ditentukan menggunakan pendekatan F1-Score Macro Average. Dalam pendekatan ini, nilai F1-Score dihitung terlebih dahulu secara terpisah untuk masing-masing kelas (F1 untuk HOTS dan F1 untuk LOTS). Setelah itu, kedua nilai tersebut dirata-rata secara aritmatika tanpa memperhitungkan jumlah sampel per kelas. Penggunaan rata-rata makro ini memastikan bahwa kemampuan model dalam mengenali kelas minoritas (HOTS) dinilai setara pentingnya dengan kelas mayoritas (LOTS), sehingga hasil evaluasi menjadi lebih adil dan representatif terhadap tujuan penelitian.

III. Hasil dan Pembahasan

A. Evaluasi Validitas Data Sintesis

Eksperimen penelitian diawali dengan evaluasi mendalam terhadap kualitas dataset yang dihasilkan oleh AI. Meskipun instruksi awal (prompt) meminta distribusi kelas yang seimbang yaitu 250 soal HOTS dan juga 250 soal LOTS, proses validasi pakar mengungkap adanya kesenjangan kualitas yang signifikan. Sebagaimana tervisualisasi dalam matriks kebingungan (confusion matrix) pada Gambar 2, teridentifikasi bahwa AI memiliki kecenderungan kuat melakukan penilaian berlebih (overclaiming) terhadap kompleksitas soal buatannya sendiri.

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

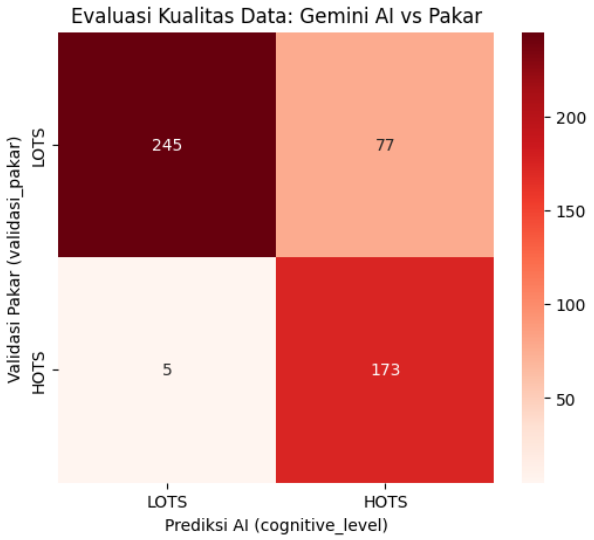


Figure 2. *Confusion Matrix Perbandingan Label AI vs Validasi Pakar*

Analisis terhadap Gambar 2 menunjukkan bahwa kesepakatan antara AI dan pakar terjadi pada 245 butir soal kategori LOTS dan 173 butir soal kategori HOTS. Namun, ketidaksesuaian yang mencolok terlihat pada tingginya angka False Positive dari sisi AI, di mana terdapat 77 butir soal yang diklaim AI sebagai kategori HOTS namun dinilai pakar hanya memenuhi kriteria LOTS. Sebaliknya, kasus underclaim terjadi dalam frekuensi yang sangat minim (5 butir). Fenomena halusinasi kategori ini menyebabkan perubahan drastis pada distribusi data final menjadi 178 soal HOTS (35,6%) dan 322 soal LOTS (64,4%). Temuan ini mengonfirmasi bahwa label otomatis dari LLM belum sepenuhnya dapat diandalkan sebagai Ground Truth tanpa verifikasi manusia. Ketidakseimbangan distribusi akibat bias penilaian ini kemudian ditangani secara teknis pada tahap klasifikasi menggunakan skema Class Weighting.

B. Komparasi Performa Model Klasifikasi

Pengujian performa dilakukan menggunakan metode 5-Fold Cross Validation untuk meminimalkan bias varians. Hasil pengujian menunjukkan perbandingan efektivitas antara representasi fitur statistik (TF-IDF) dan semantik (IndoBERT) dalam memetakan soal ke dalam kategori kognitif yang tepat. Ringkasan hasil rata-rata metrik evaluasi dari kedua skenario ditampilkan pada Tabel 3.

Metrik Evaluasi (Macro Avg)	TF-IDF + SVM	IndoBERT + SVM
Precision	74,78%	74,42%
Recall	75,97%	74,46%
F1-Score	75,13%	74,22%

Table 3. *Hasil Rata-Rata Metrik Evaluasi dari Kedua Skenario*

Berdasarkan data pada Tabel 3, Skenario 1 yang menggunakan fitur TF-IDF mencatatkan performa yang sedikit lebih unggul dibandingkan Skenario 2 yang menggunakan IndoBERT pada seluruh komponen metrik evaluasi, dengan selisih skor akhir F1-Score sebesar 0.91%. Meskipun TF-IDF menunjukkan performa lebih tinggi, perbedaan yang relatif kecil ini menunjukkan bahwa kedua pendekatan memiliki kinerja yang sebanding. Keunggulan paling signifikan dari model TF-IDF terlihat pada metrik Recall yang mencapai angka 75,97%, lebih tinggi 1,51% dibandingkan IndoBERT. Tingginya nilai Recall ini mengindikasikan bahwa fitur statistik TF-IDF yang dikombinasikan dengan proses stemming ternyata lebih sensitif dan efektif dalam menjaring kembali soal-soal HOTS yang ada di dalam dataset dibandingkan representasi embedding dari IndoBERT. Temuan kuantitatif ini cukup menarik karena menolak asumsi awal bahwa model bahasa besar atau Large Language Model akan selalu mendominasi model tradisional pada setiap kondisi eksperimen. Untuk menganalisis pola kesalahan dan keberhasilan prediksi pada tingkat kelas secara rinci, Aggregated Confusion Matrix dari masing-masing skenario diperlihatkan pada Gambar 4 dan Gambar 5 berikut.

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

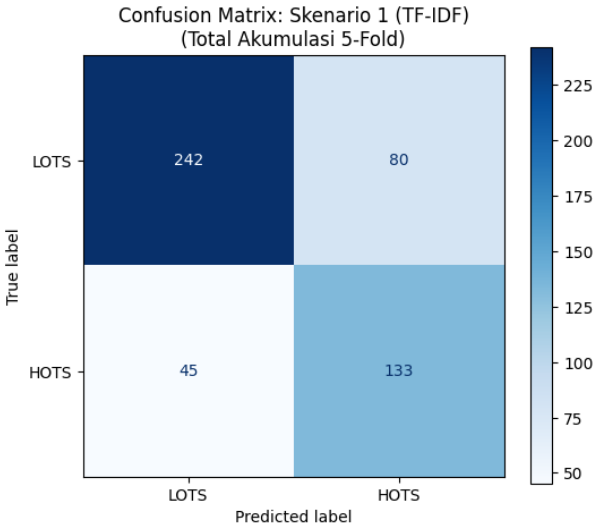


Figure 4. *Aggregated Confusion Matrix Skenario 1 (TF-IDF + SVM)*

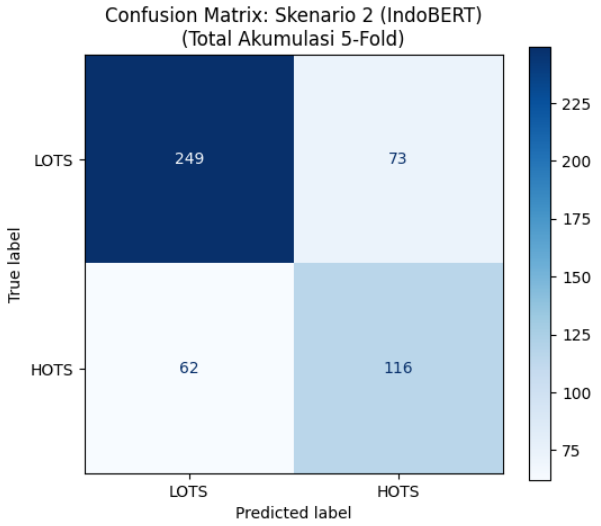


Figure 5. *Aggregated Confusion Matrix Skenario 2 (IndoBERT + SVM)*

C. Efektivitas Pendekatan Leksikal pada Data Sintesis

Pemahaman yang lebih mendalam mengenai karakteristik kesalahan prediksi diperoleh melalui analisis terhadap Aggregated Confusion Matrix (Gambar 4 dan 5) yang mengakumulasi hasil prediksi dari kelima putaran pengujian. Pada Skenario 1 yang berbasis TF-IDF, model berhasil memprediksi dengan benar 133 soal HOTS (True Positive) dan 242 soal LOTS (True Negative). Karakteristik utama dari model ini adalah agresivitasnya dalam mengenali ciri soal HOTS, yang dibuktikan dengan jumlah kesalahan False Negative atau soal HOTS yang dianggap LOTS yang relatif rendah, yaitu sebanyak 45 butir. Namun, tingginya sensitivitas ini harus dibayar dengan tingginya angka False Positive sebanyak 80 butir, di mana model cenderung terkecoh oleh soal-soal LOTS yang mengandung istilah teknis jarang muncul sehingga salah diklasifikasikan sebagai HOTS.

Sebaliknya, pada Skenario 2 yang berbasis IndoBERT, model menunjukkan karakteristik yang lebih konservatif atau hati-hati. Model ini berhasil mengenali kelas mayoritas atau LOTS dengan lebih baik, terlihat dari jumlah True Negative yang mencapai 249 butir dan angka False Positive yang lebih rendah yakni 73 butir. Kendati demikian, kehati-hatian model IndoBERT dalam memprediksi kelas positif justru berdampak pada tingginya angka False Negative yang mencapai 62 butir. Hal ini menunjukkan bahwa terdapat cukup banyak soal HOTS yang gagal ditangkap nuansa semantiknya oleh IndoBERT, sehingga soal-soal tersebut luput dari deteksi dan dianggap sebagai soal LOTS biasa.

D. Pembahasan

Untuk memberikan bukti empiris mengenai pola kesalahan, dilakukan Error Analysis pada butir soal yang diprediksi salah secara konsisten oleh kedua model (False Negative dan False Positive). Contoh spesifik kesalahan pelabelan disajikan pada Tabel 4.

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

Teks Soal	Label Pakar	Prediksi Model	Tipe Kesalahan
Anda diminta membuat sebuah fungsi rekursif untuk mencari jalur terpendek dari kota A ke kota B menggunakan peta digital. Apa 'base case' yang paling tepat untuk fungsi rekursif tersebut?	HOTS	LOTS	False Negative (FN) pada IndoBERT
Mengapa stack sangat penting dalam implementasi fungsi rekursif?	LOTS	HOTS	False Positive (FP) pada TF-IDF

Table 4. *Contoh Analisis Kesalahan Prediksi pada Data Uji*

Hasil eksperimen keseluruhan menunjukkan fenomena di mana pendekatan tradisional TF-IDF mampu mengungguli pendekatan Deep Learning IndoBERT dengan margin tipis pada kasus klasifikasi soal dengan dataset terbatas. Beberapa faktor analisis mendasar yang mempengaruhi temuan ini berkaitan dengan efisiensi data, peran pra-pemrosesan, dan arsitektur model.

Faktor pertama adalah efektivitas model pada dataset berskala kecil (Small Data Efficiency). IndoBERT merupakan model dengan jutaan parameter yang membutuhkan data berskala besar untuk mencapai optimalisasi bobot. Dengan jumlah data penelitian yang hanya berjumlah 500 butir soal, representasi vektor padat (dense vector) yang dihasilkan IndoBERT belum dapat dimanfaatkan secara maksimal oleh classifier SVM untuk membentuk batas keputusan yang presisi. Di sisi lain, TF-IDF bekerja sangat efisien pada data kecil karena secara langsung memetakan bobot kata kunci eksplisit yang menjadi pembeda utama antara soal HOTS dan LOTS, seperti keberadaan kata kerja operasional "analisis" atau "bandingkan" yang sering muncul secara harfiah.

Faktor kedua yang berkontribusi pada keunggulan Recall TF-IDF adalah peran krusial dari proses Stemming Nazief & Adriani. Dalam sekanrio TF-IDF, variasi morfologi kata seperti “menganalisis”, “dianalisis”, dan “analisa” diseragamkan menjadi satu kata kasar dasar “analisis”. Hal ini membuat dimensi fitur menjadi lebih padat dan focus sehingga sinyal fitur menjadi kuat. Berbeda halnya dengan IndoBERT yang menggunakan Tokenizer WordPiece tanpa stemming, di mana variasi kata tersebut mungkin dianggap sebagai token yang berbeda-beda. Pada dataset yang kecil, perpecahan token ini menyebabkan sinyal fitur menjadi terpecah atau diluted, sehingga model kesulitan menangkap pola yang konsisten untuk kelas HOTS.0 Sejalan dengan hasil analisis tersebut, keunggulan pendekatan TF-IDF yang ditunjukkan dalam penelitian ini bersifat kontekstual, yaitu terbatas pada kondisi dataset berukuran kecil dan domain spesifik algoritma pemrograman, sehingga tidak dapat digeneralisasi secara langsung pada skenario data besar atau lintas domain tanpa pengujian lanjutan.

Faktor ketiga adalah keterbatasan pengguna IndoBERT sebagai frozen feature extractor. Dalam penelitian ini, model IndoBERT digunakan hanya untuk mengekstraksi fitur tanpa melakukan fine tuning atau pelatihan ualng bobot pada domain spesifik. Kondisi ini menyebabkan model hanya mengandalkan pengetahuan umum bahasa Indonesia yang dilikinya sejak pra-pelatihan, tanpa beradaptasi secara mendalam terhadap istilah teknik spesifik pada domain “Algoritma dan Pemrorgaman”. Akibatnya, model menjadi kurang sensitif terhadap nuansa kalimat HOTS yang spesifik pada domain ini, yang tercermin dari tingginya angka False Negative pada skenario kedua.

Secara keseluruhan, penelitian ini menyimpulkan bahwa meskipun IndoBERT menawarkan kemampuan pemahaman konteks yang kaya, pada kondisi dataset yang sangat terbatas dan spesifik domain, pendekatan statistik TF-IDF yang dilengkapi dengan pra-pemrosesan stemming yang matang terbukti menjadi solusi yang lebih tangguh, efisien secara komputasi, dan memiliki sensitivitas yang lebih baik dalam mendeteksi soal HOTS.

Berdasarkan temuan tersebut, arah penelitian selanjutnya perlu difokuskan pada peningkatan adaptasi model berbasis transformer terhadap domain soal algoritma, terutama melalui penerapan fine-tuning. Selain itu keterbatasan data dapat diatasi dengan eksplorasi teknik data augmentation atau synthetic data generation berbasis LLM. Pengembangan arsitektur hibrida yang mengombinasikan fitur leksikal TF-IDF dan representasi semantic BERT juga berpotensi meningkatkan ketepatan klasifikasi. Terakhir, perluasan domain dan variasi jenis soal menjadi langkah penting untuk menguji daya generalisasi temuan penelitian ini.

IV. Kesimpulan

Mengacu pada serangkaian eksperimen dan analisis yang telah diselesaikan, studi ini menyimpulkan bahwa pendekatan statistik tradisional Term Frequency-Inverse Document Frequency (TF-IDF), yang dipadukan dengan algoritma Support Vector Machine (SVM) terbukti lebih efektif dibandingkan pendekatan semantik berbasis Deep Learning (IndoBERT) dalam klasifikasi tingkat kognitif soal "Algoritma dan Pemrograman" pada konteks dataset yang terbatas. Model TF-IDF mencatatkan keunggulan performa pada seluruh metrik evaluasi utama, dengan pencapaian F1-Score Macro sebesar 75,13% dan Recall sebesar 75,97%.

Keunggulan ini menegaskan bahwa untuk dataset berskala kecil (500 butir data) dengan ketidakseimbangan kelas, metode

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

representasi fitur yang sederhana namun terarah mampu memberikan hasil yang lebih sensitif dalam mendeteksi soal kategori HOTS dibandingkan model bahasa besar yang kompleks namun statis. Keberhasilan skenario TF-IDF didorong oleh efisiensi pemetaan kata kunci eksplisit yang diperkuat oleh proses stemming Nazief & Adriani. Penyeragaman variasi morfologi kata menjadi bentuk dasar terbukti ampuh memadatkan dimensi fitur, sehingga sinyal pembeda antara kelas HOTS dan LOTS menjadi lebih tegas bagi algoritma klasifikasi. Sebaliknya, performa IndoBERT terkendala oleh arsitekturnya yang membutuhkan data besar untuk optimalisasi. Penggunaan mekanisme frozen feature extractor tanpa fine tuning menyebabkan model gagal beradaptasi dengan terminology spesifik domain, yang mengakibatkan tingginya angka False Negative atau kegagalan model dalam mengenali nuansa implisit pada soal HOTS. Secara keseluruhan, temuan ini menunjukkan bahwa mode Large Languange Model (LLM) tidak perlu mendominasi performa model konvensional, khususnya pada konteks dataset kecil dan domain spesifik algoritma pemrograman.

Implikasi praktis dari temuan ini merekomendasikan penggunaan pendekatan leksikal seperti TF-IDF sebagai solusi yang lebih robust dan efisien secara komputasi untuk evaluasi soal otomatis pada lingkungan dengan ketersediaan data yang minim, sementara penggunaan model berbasis transformer memerlukan strategi pelatihan lanjutan atau jumlah data yang jauh lebih masif untuk dapat melampaui performa metode statistik. Adapun batasan penelitian ini mencakup keterbatasan ukuran dataset, ruang lingkup materi yang hanya berfokus pada algoritma dan pemrograman, serta tidak dilakukannya proses fine-tuning pada model IndoBERT, sehingga hasil penelitian belum dapat digeneralisasi pada domain lain atau skala data yang lebih besar. Oleh karena itu, penelitian selanjutnya disarankan untuk menguji skema pelatihan lanjutan dan perluasan domain guna memperoleh hasil yang lebih komprehensif.

Ucapan Terima Kasih

Penulis menyampaikan terima kasih atas kontribusi serta dukungan yang telah diterima sepanjang proses penelitian ini. Ucapan terima kasih secara mendalam ditujukan kepada para dosen ahli yang telah meluangkan waktu dan keahlian untuk memvalidasi dataset soal, di mana penilaian ahli ini menjadi kunci penting dalam menentukan Ground Truth penelitian. Apresiasi juga diberikan kepada rekan-rekan sekelas dan teman-teman satu bimbingan yang senantiasa memberikan semangat, dorongan moral, dan berbagi wawasan selama perjalanan akademik. Dukungan kolektif ini merupakan faktor pendorong utama yang memungkinkan penelitian ini berhasil diselesaikan.

References

1. A. Ahmed, E. Kerr, and A. O'Malley, "Quality assurance and validity of AI-generated single best answer questions," *BMC Medical Education*, vol. 25, no. 1, p. 300, Feb. 2025, doi: 10.1186/s12909-025-06881-w.
2. A. A. Alsagoafi and H. S. Alomran, "Revolutionizing Assessment: Leveraging ChatGPT for Automated Item Generation: An AI Driven Exploratory Study with EFL Teachers," *World Journal of English Language*, vol. 15, no. 6, p. 385, July 2025, doi: 10.5430/wjel.v15n6p385.
3. G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.
4. Y. Susanti, T. Tokunaga, and H. Nishikawa, "Integrating automatic question generation with computerised adaptive test," *Research and Practice in Technology Enhanced Learning*, vol. 15, no. 1, p. 9, Dec. 2020, doi: 10.1186/s41039-020-00132-w.
5. L. O. Wilson and C. Leslie, "Anderson and Krathwohl Bloom's Taxonomy Revised."
6. K. U. Danyaro, S. Abdullahi, A. S. Abdallah, and H. Chiroma, "Hallucinations in Large Language Models for Education: Challenges and Mitigation," *International Journal of Teaching and Learning in Education*, vol. 4, no. 6, pp. 13–19, 2025, doi: 10.22161/ijtle.4.6.2.
7. N. E. Mustaffa, K. E. Lai, C. Preece, and F. Y. Wong, "A Bibliometric Review of Large Language Model Hallucination," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5065151.
8. M. O. Omopekunola and E. Y. Kardanova, "Automatic generation of physics items with Large Language Models (LLMs)," *REID Research and Evaluation in Education*, vol. 10, no. 2, pp. 168–185, Oct. 2024, doi: 10.21831/reid.v10i2.76864.
9. A. Herrmann-Werner et al., "Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study," *Journal of Medical Internet Research*, vol. 26, p. e52113, Jan. 2024, doi: 10.2196/52113.
10. P. Sheridan, Z. Ahmed, and A. A. Farooque, "A Fisher's Exact Test Justification of the TF–IDF Term-Weighting Scheme," *The American Statistician*, pp. 1–11, Sept. 2025, doi: 10.1080/00031305.2025.2539241.
11. Z. Xu, M. Chen, K. Q. Weinberger, and F. Sha, "An alternative text representation to TF-IDF and Bag-of-Words," Jan. 28, 2013, arXiv: arXiv:1301.6770, doi: 10.48550/arXiv.1301.6770.
12. B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Oct. 08, 2020, arXiv: arXiv:2009.05387, doi: 10.48550/arXiv.2009.05387.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805, doi: 10.48550/arXiv.1810.04805.
14. M. Selvam and R. González Vallejo, "Human-in-the-Loop Models for Ethical AI Grading: Combining AI Speed with Human Ethical Oversight," *EthAIca*, vol. 4, p. 413, Aug. 2025, doi: 10.56294/ai2025413.
15. C. Petridis, "Text Classification: Neural Networks VS Machine Learning Models VS Pre-trained Models," Dec. 30, 2024, arXiv: arXiv:2412.21022, doi: 10.48550/arXiv.2412.21022.
16. S. Chanda and S. Pal, "The Effect of Stopword Removal on Information Retrieval for Code-Mixed Data Obtained Via Social Media," *SN Computer Science*, vol. 4, no. 5, p. 494, June 2023, doi: 10.1007/s42979-023-01942-7.
17. Arif Bijaksana Putra Negara, "The Influence Of Applying Stopword Removal And Smote On Indonesian Sentiment Classification," *Lontar Komputer Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 03, pp. 172–185, Oct. 2025, doi: 10.24843/LKJITI.2023.v14.i03.p05.
18. X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece Tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 2089–2103, doi: 10.18653/v1/2021.emnlp-main.160.
19. A. Nayak, H. Timmapathini, K. Ponnalagu, and V. Gopalan Venkoparao, "Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words," in *Proceedings of the First Workshop on Insights from Negative Results in NLP*, Association for Computational Linguistics, 2020, pp. 1–5, doi: 10.18653/v1/2020.insights-1.1.
20. Y. Wahba, N. Madhavji, and J. Steinbacher, "A Comparison of SVM Against Pre-trained Language Models (PLMs) for Text Classification Tasks," in *Machine Learning, Optimization, and Data Science*, Lecture Notes in Computer Science, vol. 13811, Springer Nature Switzerland, 2023, pp. 304–313, doi: 10.1007/978-3-031-25891-6_23.

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1881

21. B. Zhu, X. Jing, L. Qiu, and R. Li, "An Imbalanced Data Classification Method Based on Hybrid Resampling and Fine Cost Sensitive Support Vector Machine," *Computers, Materials & Continua*, vol. 79, no. 3, pp. 3977–3999, 2024, doi: 10.32604/cmc.2024.048062.
22. L. Hakim, A. Sobri, L. Sunardi, and D. Nurdiansyah, "Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn," *Jurnal Digital Teknologi Informasi*, vol. 7, no. 2, p. 14, Feb. 2025, doi: 10.32502/digital.v7i2.9429.

ISSN 2598-9936 (online), https://ijins.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo