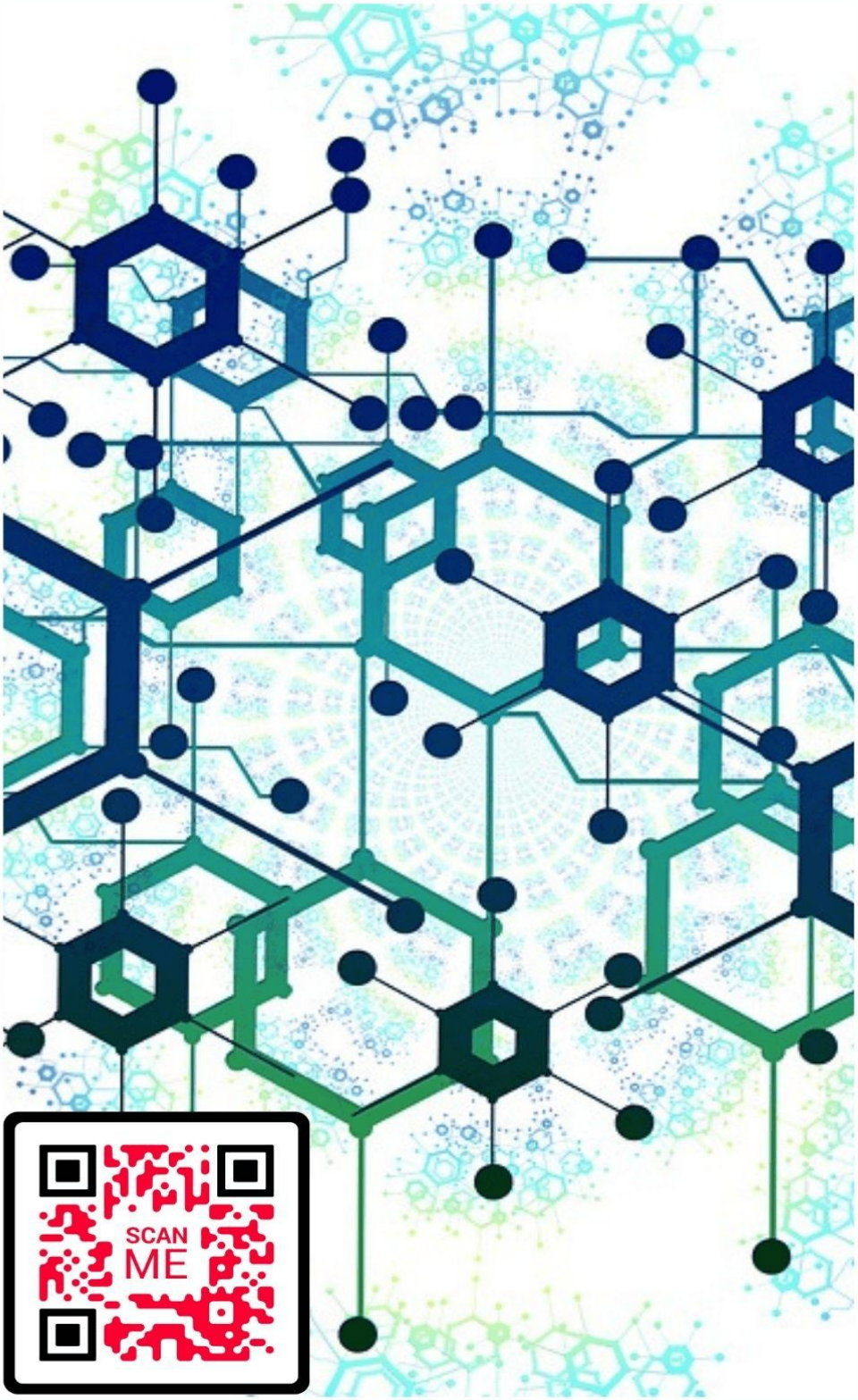
Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

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.v27i1.1880

Table Of Contents

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

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.v27i1.1880

EDITORIAL TEAM

Indonesian Journal of Innovation Studies

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

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.v27i1.1880

Automatic Classification of Artificial Intelligence Generated Question Difficulty Levels

Klasifikasi Otomatis Tingkat Kesulitan Soal Hasil Kecerdasan Buatan

Vina Najahah, vina.najahah.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: Determining question difficulty is a fundamental requirement in educational assessment to support valid evaluation and systematic question curation. **Specific Background:** The increasing use of artificial intelligence for automatic question generation produces large volumes of linguistically diverse items, making manual difficulty labeling time-consuming and subjective. **Knowledge Gap:** Despite extensive research on text-based difficulty prediction, lightweight and reproducible pipelines for multi-level difficulty classification of AI-generated questions remain limited. **Aims:** This study aims to develop and evaluate an automatic classification pipeline for three difficulty levels of AI-generated multiple-choice questions using TF-IDF text representation and a Random Forest classifier. **Results:** The proposed pipeline achieved a test accuracy of 70.98%, exceeding the random guessing baseline, with the highest F1-score observed in the easy class (78.45%) and the lowest in the medium class (65.32%), indicating greater ambiguity in intermediate difficulty questions. **Novelty:** This study presents a reproducible and interpretable classification workflow specifically applied to expert-labeled AI-generated questions with high inter-rater reliability. **Implications:** The findings support the use of lexical feature–based classification as an initial pre-curation and difficulty filtering tool in AI-assisted educational assessment systems.

Highlights

- The classification pipeline distinguishes three difficulty levels using only textual features
- Medium difficulty questions exhibit the highest classification ambiguity
- Lexical patterns contribute consistently to difficulty level separation

Keywords

Question Difficulty Classification; AI Generated Questions; TF-IDF Representation; Random Forest Classifier; Educational Assessment

Published date: 2026-01-06

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

I. Pendahuluan

Penentuan tingkat kesulitan soal merupakan aspek fundamental dalam sistem evaluasi pembelajaran karena berpengaruh langsung terhadap validitas asesmen, penyesuaian materi ajar, serta pengembangan sistem pembelajaran adaptif. Secara tradisional, tingkat kesulitan soal ditentukan melalui analisis pakar atau berdasarkan data respons siswa, namun pendekatan tersebut memerlukan waktu, biaya, dan sumber daya yang besar serta berpotensi mengandung subjektivitas penilai. Tantangan ini semakin meningkat seiring dengan pemanfaatan Artificial Intelligence (AI), khususnya Large Language Models (LLM), yang mampu menghasilkan soal secara otomatis dalam jumlah besar dan dengan variasi linguistik yang luas. Meskipun model-model ini menunjukkan potensi besar dalam pembuatan soal edukatif, berbagai studi menyoroti perbedaan performa antar-model dalam hal relevansi, tingkat kesulitan, serta kesesuaian pedagogis, sehingga proses kurasi dan validasi kualitas soal menjadi semakin penting [1], [2]. Dalam konteks ini, proses kurasi dan validasi tingkat kesuslitan soal menjadi kbutuhan yang mendesak.

Perkembangan bidang Natural Languange Processing (NLP) membuka peluang untuk melakukan prediksi tingkat kesulitan soal secara otomatis hanya berdasarkan karakteristik linguistic teks, tanpa bergantung pada data respons siswa. Kajian sistematis terbaru menunjukkan bahwa pendekatan berbasis teks telah banyak digunakan untuk memprediksi tingkat kesulitan item asesmen, dengan memanfaatkan fitur linguistic seperti ditribusi kata, kompleksitas sintakis,dan variasi leksikal [2]. Pendekatan ini dinilai sangat relevan untuk konteks soal yang dihasilkan oleh AI, karena pada tahap awal penggunaan soal umumnya belum tersedia data empiris respons siswa [3]. Oleh karena itu, diperlukan representasi teks yang efisien dan informatif untuk mendukung proses klasifikasi tingkat kesulitan secara otomatis.

Salah satu teknik representasi teks yang paling umum dan terbukti efektif dalam NLP adalah Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF merepresentasikan teks ke dalam bentuk numerik dengan mempertimbangkan frekuensi kemunculan kata dan tingkat kepentingannya dalam keseluruhan dokumen, sehingga mampu menangkap karakteristik linguistik yang relevan. Penelitian sebelumnya menunjukkan bahwa TF-IDF efektif digunakan dalam berbagai tugas klasifikasi teks, termasuk klasifikasi sentimen dan teks pendidikan, serta mampu memberikan kinerja yang stabil pada berbagai domain [4]. Selain itu, TF-IDF sering digunakan sebagai baseline yang kuat karena sifatnya yang ringan, mudah diimplementasikan, dan mudah direproduksi dalam penelitian klasifikasi teks. Studi pada teks pendidikan menunjukkan bahwa TF-IDF berfungsi sebagai baseline awal yang stabil untuk klasifikasi otomatis pada berbagai tingkatan kurikulum sekolah, mencapai performa tinggi dengan model klasik seperti linear regression dan cosine similiarty [5]. Dengan demikian, TF-IDF menjadi fondasi penting dalam pengembangan pipeline klasifikasi tingkat kesulitan soal berbasis teks.

Dalam hal pemodelan, algoritma machine learning berbasis decision tree, khususnya Rmadom Forest, telah banyak diterapkan dalam berbagai tugas klasifikasi teks dan domain pendidikan. Ramdom Forest memiliki keunggulan dalam menangani data berdimensi tinggi, mengurangi risiko overfitting, serta memberikan performa yang stabil pada dataset dengan distribusi fitur yang kompleks. Meskipun diperkenalkan sebagai metode dasar, Ramdom Forest tetap relevan dan banyak digunakan dalam penelitian modern baik sebagai baseline maupun model utama untuk klasifikasi teks. Kombinasi antara stabilitas, kemampuan generalisasi, dan interpretabilitas menjadikan metode ini pilihan populer dalam berbagai studi NLP kontemporer [6]. Studi terkini menunjukkan bahwa kombinasi TF-IDF dan Random Forest mampu menghasilkan performa yang kompetitif dalam tugas klasifikasi tingkat kesulitan soal dan analisis asesmen pendidikan [7], serta tetap efektif ketika dikombinasikan dengan teknik penyeimbangan data seperti Synthetic Minority Over-sampling Technique (SMOTE) [8]. Berdasarkan hal ini, integrasi antara representasi TF-IDF dan Random Forest berpotensi menghasilkan sistem klasifikasi yang akurat dan efisien.

Meskipun berbagai penelitian telah membahas prediksi kesulitan soal dan klasifikasi teks pendidikan, masih terdapat sejumlah tantangan yang belum sepenuhnya teratasi. Soal yang dihasilkan oleh LLM memiliki variasi bahasa, struktur kalimat, dan gaya penyajian yang lebih beragam dibandingkan soal konvensional, sehingga meningkatkan kompleksitas dalam proses klasifikasi otomatis [1], [9]. Selain itu, ketersediaan dataset berlabel tingkat kesulitan yang reliabel masih terbatas, dan kualitas ground truth sangat bergantung pada konsistensi penilaian pakar. Oleh karena itu, Cohen's Kappa tetap menjadi metrik utama untuk mengukur konsistensi antar-penilai, terutama dalam anotasi teks edukatif, langsung sejalan dengan konteks validasi label di penelitian klasifikasi tingkat kesulitan soal [10]. Kajian sistematis terbaru juga menekankan bahwa masih diperlukan pipeline klasifikasi yang ringan, terstandar, dan mudah direproduksi untuk tugas prediksi tingkat kesulitan item berbasis teks [2]. Untuk mengatasi kendala tersebut, diperlukan pipeline yang tidak hanya akurat, tetapi juga ringan dan mudah direplikasi untuk aplikasi praktis.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi suatu pipeline klasifikasi otomatis tiga tingkat kesulitan pada soal yang dihasilkan oleh AI, yaitu mudah, sedang, dan sulit, dengan memanfaatkan representasi TF-IDF dan algoritma Random Forest. Penelitian ini juga mempertimbangkan penanganan ketidakseimbangan kelas melalui teknik data balancing seperti SMOTE apabila diperlukan. Evaluasi model dilakukan dengan mengukur performa klasifikasi pada data uji, kestabilan model melalui k-fold cross validation, serta validasi kualitas label menggunakan koefisien Cohen's Kappa sebagai ukuran reliabilitas antar-penilai. Kontribusi utama penelitian ini meliputi penyusunan workflow klasifikasi yang reproducible dan aplikatif, yang mencakup tahapan pengolahan data, praproses teks, esktraksi fitur, penyeimbangan data, pelatihan model, dan evaluasi performa. Selain itu, penelitian ini menekankan pentingnya analisis reliabilitas gorund truth dalam klasifikasi tingkat kesulitan soal.

Secara ringkas, penelitian ini menggunakan representasi TF-IDF dengan n-gram terbatas (1-2) untuk menjaga kestabilan fitur, dikombinasikan dengan model Ramdom Forest yang dikonfigurasi secara seimbang. Dataset dibagi secara stratified dengan rasio 80:20 antara data latih dan data uji, serta dievaluasi menggunakan 5-fold crocss- validation. Teknik SMOTE diterapkan apabila distribusi kelas menunjukkan ketimpangan yang signifikan. Penelitian ini dibatasi pada klasifikasi tiga

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

tingkat kesulitan berbasis teks dan tidak membahas penerapan model deep learning seperti BERT atau transformer lainnya. Fokus utama diarahkan pada pendekatan klasik yang ringan, terukur, dan mudah direplikasi.

Oleh karena itu, penelitian ini berfokus pada prediksi intended difficulty atau expert-perceived difficulty, yaitu tingkat kesulitan yang dipersepsikan oleh pakar berdasarkan karakteristik kognitif dan linguistik soal. Pendekatan ini berbeda dari kesulitan empiris yang umumnya diukur menggunakan data respons peserta tes (seperti dalam Item Response Theory atau Classical Test Theory), yang belum tersedia pada tahap awal kurasi soal AI-generated. Dengan demikian, klasifikasi yang dikembangkan bertujuan untuk mendukung proses pra-validasi dan kurasi awal sebelum soal diujikan kepada siswa.

Studi Literatur

Penelitian mengenai klasifikasi tingkat kesulitan soal berbasis teks telah banyak dilakukan dengan memanfaatkan pendekatan NLP dan ML. Salah satu penelitian mengusulkan kombinasi TF-IDF dan Random Forest untuk mengklasifikasikan tingkat kesulitan soal pilihan ganda, dan hasilnya menunjukkan akurasi tinggi pada soal buatan manusia [7]. Namun, penelitian tersebut masih terbatas pada soal manual dan belum menguji efektivitas metode pada soal yang dihasilkan oleh kecerdasan buatan (AI-generated), yang memiliki variasi linguistik berbeda.

Studi lain menggunakan algoritma ML untuk mengklasifikasikan soal matematika berdasarkan kompleksitas dan struktur pertanyaannya [11]. Pendekatan ini efektif dalam mengidentifikasi tingkat kesulitan, tetapi cakupannya terbatas pada satu bidang dan belum diuji pada soal lintas subjek atau soal generatif. Penelitian lanjutan membandingkan beberapa algoritma ML seperti Decision Tree, Naïve Bayes, dan Random Forest untuk klasifikasi tingkat kesulitan soal matematika [12]. Meskipun Random Forest memberikan hasil terbaik, analisis masih berfokus pada akurasi numerik tanpa memperhatikan fitur linguistik atau interpretabilitas model.

Dalam konteks soal generatif, beberapa penelitian mulai mengevaluasi kualitas AI-generated questions. Sebuah studi berskala besar menilai kualitas ujian yang dihasilkan oleh model AI dan menemukan bahwa meskipun hasilnya potensial, klasifikasi tingkat kesulitan belum dikaji secara mendalam [13]. Penelitian lain menilai kesesuaian soal AI terhadap Bloom's Taxonomy dan menyimpulkan bahwa integrasi metode NLP dan ML untuk klasifikasi otomatis masih jarang dilakukan [14].

Kajian sistematis lainnya menunjukkan bahwa sebagian besar penelitian prediksi tingkat kesulitan soal berbasis teks masih menggunakan klasifikasi biner dan belum mengakomodasi klasifikasi multi-level [15]. Dalam studi komparatif terbaru, pendekatan berbasis supervised learning yang memanfaatkan ketidakpastian prediksi dari model state-of-the-art LLM bahkan mengungguli prediksi dosen pada tugas menilai kesulitan soal True/False dalam konteks Neural Networks dan Machine Learning [16]. Meskipun demikian, temuan ini masih terbatas pada domain spesifik soal, dan masih diperlukan penelitian lanjutan untuk memperluasnya ke berbagai jenis dan format soal, termasuk multi-pilihan atau soal generatif.

Secara umum, literatur terdahulu menunjukkan bahwa pendekatan berbasis NLP dan ML memiliki potensi besar untuk klasifikasi tingkat kesulitan soal. Namun, terdapat celah penelitian (research gap) berupa kurangnya eksplorasi pada soal AI-generated dengan kompleksitas linguistik dan variasi konteks yang lebih tinggi. Oleh karena itu, penelitian ini mengusulkan model klasifikasi multi-level berbasis TF-IDF n-gram dan Random Forest, yang diharapkan dapat memberikan hasil klasifikasi yang lebih akurat dan adaptif terhadap karakteristik teks generatif.

II. Metode

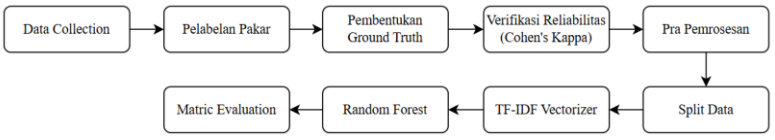


Figure 1. *Alur Penelitian*

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk mengembangkan dan mengevaluasi sistem klasifikasi otomatis tingkat kesulitan soal yang dihasilkan oleh AI. Metode yang digunakan mencakup tahapan pengumpulan data, penentuan ground truth, praproses teks, ekstraksi fitur, pembagian data, penyeimbangan kelas, pelatihan model, serta evaluasi performa. Alur metode penelitian dirancang secara sistematis dan reproducible agar hasil penelitian dapat diverifikasi dan direplikasi.

Tingkat Kesulitan	Jumlah (n)	Presentase (%)
Mudah	174	34.32
Sedang	185	36.49
Sulit	148	29.19
Total	507	100

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

Table 1. *Distribusi Label*

Dataset yang digunakan dalam penelitian ini berupa kumpulan soal berbasis teks yang dihasilkan oleh sistem AI, dengan total sebanyak 507 sampel. Sebagaimana ditunjukkan pada Tabel 1, setiap soal diklasifikasikan ke dalam tiga tingkat kesulitan, yaitu Mudah, Sedang, dan Sulit, dengan distribusi masing-masing sebesar 174 data (34.32%), 185 data (36.49%), dan 148 data (29.19%). Penentuan label tingkat kesulitan dilakukan oleh dua orang pakar secara independen dengan mengacu pada kriteria pelabelan terstruktur yang disusun berdasarkan taknonomi kognitif pembelajaran serta karakteristik soal pemrograman. Setiap pakar memberikan label sesuai dengan indikator kognitif, komoleksitas penalaran, dan karakteristik kode atau algoritma yang terlibat, sehingga proses anotasi bersifat sistematis, konsisten, dan tidak bergantung pada subjektivitas semata.

Untuk mengevaluasi reliabilitas antar-pakar pada proses anotasi data, penelitian ini menggunakan koefisien Cohen's Kappa, yang dirancang untuk mengukur tingkat kesepakatan dua penilai dengan mempertimbangkan kemungkinan kesepakatan yang terjadi secara kebetulan. Penggunaan metric ini bertujuan untuk memastikan bahwa label yang dihasilkan memiliki tingkat konsistensi dan objektivitas yang memadai sebelum digunakan sebagai gorund truth dalam pengembangan model klasifikasi. Statistik ini telah digunakan secara luas dalam studi anotasi teks untuk memastikan bahwa label yang dihasilkan memiliki tingkat konsistensi yang memadai, sehingga dapat diandalkan untuk pelatihan dan evaluasi model otomatis [17]. Secara matematis, koefisien Cohen's Kappa (κ) dirumuskan sebagai berikut:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

Dengan $P_o$sebagai proporsi kesepakatan yang teramati (observed agreement) dan $P_e$sebagai proporsi kesepakatan yang diharapkan terjadi secara kebetulan (expected agreement). Cohen's Kappa dirancang untuk mengukur tingkat kesepakatan antar-penilai yang melampaui kemungkinan kesepakatan acak (beyond chance) serta memberikan ukuran reliabilitas yang terstandarisasi antara dua penilai, sebagaimana dijelaskan dalam penelitian [18]. Selain itu, metrik ini banyak digunakan dalam evaluasi inter-rater reliability karena kemampuannya dalam mengoreksi bias yang timbul akibat kesepakatan acak [19]. Berdasarkan formulasi tersebut, nilai Cohen's Kappa kemudian dihitung untuk menilai tingkat reliabilitas anotasi antar-pakar pada data yang digunakan dalam penelitian ini.

Metrik	Nilai
Cohen's Kappa	0.9733
Interpretasi	Almost perfect agreement
Persentase Konsensus	98.22%
Jumlah Konsensus	498
Total Data	507

Tabel 2. *Reliabilitas Pakar*

Nilai reliabilitas anotasi antar-pakar pada Tabel 2. menunjukkan kualitas ground truth yang sangat tinggi. Cohen's Kappa sebesar 0.9733 mengindikasikan tingkat kesepakatan almost perfect agreement, yang berarti kesesuaian label antar-pakar berada jauh di atas kemungkinan kesepakatan secara kebetulan. Hal ini diperkuat oleh persentase konsensus sebesar 98.22%, dengan 498 dari 507 data memiliki label yang identik antara kedua pakar. Tingginya nilai kesepakatan ini menjustifikasi penggunaan hasil anotasi sebagai ground truth dalam proses pelatihan dan evaluasi model klasifikasi, karena reliabilitas label terjamin dan potensi bias anotasi dapat diminimalkan.

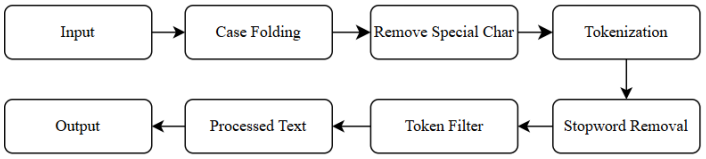


Figure 2. *Text Processing*

Tahap processing teks dilakukan untuk membersihkan dan menstandarkan data sebelum ekstraksi fitur, sebagaimana ditunjukkan pada Gambar 1. Proses dimulai dari input teks mentah yang kemudian melalui case folding untuk mengubah seluruh teks menjadi huruf kecil. Selanjutnya, dilakukan penghapusan karakter alfanumerik yang relevan. Teks yang telah dibersihkan kemudian diproses melalui tokenization untuk memecah teks menjadi token kata, diikuti dengan stopword removal menggunakan daftar stopword bahasa Indonesia dan bahasa Inggris. Tahap akhir adalah token filter, yaitu, penyaringan token berdasarkan panjang karakter, di mana hanya token dengan panjang minimal tiga karakter yang dipertahankan. Hasil dari rangkaian proses ini adalah processed text yang siap digunakan pada tahap ekstraksi fitur.

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

Tahap	Hasil
Original	Apa tujuan utama dari 'encapsulation' dalam pemrograman berorientasi objek?
Cleaned	apa tujuan utama dari encapsulation dalam pemrograman berorientasi objek
Processed	tujuan utama encapsulation pemrograman berorientasi objek
Reduksi Token	9 → 6 (33.3%)

Table 3. *Contoh Hasil Praproses Teks pada Soal dengan Label Mudah*

Tabel 3. menyajikan contoh hasil praproses teks pada satu soal dengan label Mudah. Terlihat bahwa tahapan praproses mampu mengurangi jumlah token dari 9 menjadi 6 (33.3%) tanpa menghilangkan makna utama soal, sehingga teks menjadi lebih ringkas dan representatif untuk proses ekstraksi fitur berbasis TF-IDF.

Pada tahap representasi fitur teks, penelitian ini menggunakan metode TF-IDF untuk mengubah dokumen teks menjadi vektor numerik berbobot. Metode TF-IDF banyak digunakan dalam pemrosesan bahasa alami karena mampu merepresentasikan tingkat kepentingan suatu kata dalam dokumen dengan mempertimbangkan frekuensinya secara lokal dan global dalam korpus, sebagaimana diterapkan pada penelitian terdahulu [20]. Secara matematis, bobot TF-IDF suatu kata dirumuskan sebagai berikut:

$$tf\_idf_{x,y} = tf_{x,y} \times idf\ (\omega) \quad (2)$$

Di mana $tf_{x,y}$ menyatakan term frequency kata $x$dalam dokumen y, dan $idf(\omega)$merupakan inverse document frequency dari kata tersebut. Nilai term frequency dihitung sebagai proporsi kemunculan suatu kata terhadap keseluruhan kata dalam dokumen, yang dirumuskan sebagai:

$$tf_{x,y} = \frac{n_{x,y}}{\Sigma_{i \in y} n_x(i)} \quad (3)$$

Di mana $n_{x,y}$menunjukkan jumlah kemunculan kata $x$dalam dokumen y. Formulasi ini merepresentasikan tingkat kepentingan suatu kata berdasarkan frekuensi relatifnya dalam sebuah dokumen, sehingga kata yang muncul lebih sering akan memperoleh bobot yang lebih besar dibandingkan kata yang jarang muncul. Sementara itu, inverse document frequency digunakan untuk menurunkan bobot kata-kata yang sering muncul di banyak dokumen dan dirumuskan sebagai:

$$idf_{(\omega)} = log \frac{D}{df} \quad (4)$$

Di mana D adalah jumlah total dokumen dalam korpus dan dfmenyatakan jumlah dokumen yang mengandung kata tersebut. Formulasi TF-IDF ini mengikuti pendekatan yang digunakan dalam penelitian [20], yang menekankan keseimbangan antara informasi lokal dan global dalam representasi fitur teks.

Parameter	Value	Rationale
max_features	150	Reduce dimensionality, prevent overfitting on small dataset (507 samples)
ngram_range	(1, 2)	Capture unigrams & bigrams to preserve n-gram information
min_df	3	Remove terms appearing in < 3 documents (too rare)
max_df	0.80	Remove terms appearing in > 80% documents (too common)
sublinear_tf	True	Sublinear term frequency scaling to dampen frequent terms
norm	L2	L2 normalization for text classification stability
use_idf	True	Enable IDF (inverse document frequency) weighting
smooth_idf	True	Smooth IDF weights to prevent zero divisions

Table 4. *TF-IDF Feature Extraction Parameters*

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

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

Konfigurasi TF-IDF yang digunakan sesuai dengan Tabel 4. meliputi max_features = 150, ngram_range = (1,2), min_df = 3, max_df = 0.80, sublinear_tf = True, dan norm = L2, dengan pembobotan IDF diaktifkan (use_idf = True) serta smooth_idf = True. Penggunaan unigram dan bigram bertujuan untuk menangkap konteks kata tunggal maupun pasangan kata, sedangkan pembatasan jumlah fitur dilakukan untuk mengurangi dimensi vektor dan meminimalkan risiko overfitting pada dataset berukuran relatif kecil.

Metrik	Value
Total Features	150
Total Samples (Train)	405
Feature-to-Sample Ratio	0.370
Ideal Range	0.100 - 0.300

Table 5. *Feature-to-Sample Ratio Validation*

Tabel 5. menyajikan hasil analisis rasio antara jumlah fitur dan jumlah sampel data latih. Rasio fitur terhadap sampel sebesar 0.370 menunjukkan bahwa kompleksitas fitur masih dalam batas yang dapat dikelola untuk pelatihan model, meskipun sedikit berada di atas rentang ideal, sehingga strategi pembatasan fitur tetap diperlukan untuk menjaga kestabilan dan generalisasi model. Meskipun rasio fitur terhadap sampel (0.370) sedikit berada di atas rentang ideal, risiko overfitting diminimalkan melalui strategi mitigasi seperti pembatasan kedalaman pohon (*max_depth = 12*), penggunaan class_weight = 'balanced', serta evaluasi berbasis cross-validation yang ketat. Dengan demikian, model tetap dapat diandalkan untuk generalisasi meskipun dalam kondisi rasio yang sedikit tinggi.

Dataset dibagi menjadi data latih dan data uji menggunakan metode stratified split dengan rasio 80:20. Pendekatan ini dipilih untuk memastikan bahwa proporsi kelas pada data latih dan data uji tetap mempresentasikan distribusi kelas pada keseluruhan dataset. Data latih terdiri dari N_train sampel, sedangkan data uji terdiri dari N_test sampel. Pembagian ini bertujuan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

Parameter	Value
n_estimators	200
max_depth	12
min_samples_split	10
min_samples_leaf	4
max_features	sqrt
class_weight	Balanced
random_state	42

Table 6. *Random Forest Model Parameters*

Model klasifikasi yang digunakan dalam penelitian ini adalah Random Forest Classifier dengan konfigurasi parameter sebagaimana ditunjukkan pada Tabel 6. Parameter model disesuaikan untuk menyeimbangkan kompleksitas dan kemampuan generalisasi, di mana jumlah pohon (n_estimators = 200) digunakan untuk meningkatkan stabilitas prediksi, sementara pembatasan kedalaman pohon (max_depth = 12) serta pengaturan min_samples_split = 10 dan min_samples_leaf = 4 diterapkan untuk mengurangi risiko overfitting. Penggunaan max_features = 'sqrt' bertujuan meningkatkan keragaman antar tree, sedangkan class_weight = 'balanced' digunakan untuk menangani ketidakseimbangan kelas. Selain itu, penerapan boostrap = True dan ramdom_state= 42 memastikan proses pelatihan bersifat konsisten dan dapat direproduksi.

Evaluasi model dilakukan menggunakan 5-fold stratified cross –validation padda keseluruhan dataset untuk mengukur kestabilan performa model. Metric evaluasi yang digunakan mencakup precision, recall, dan F1-score untuk setiap kelas, serta nilai support sebagai jumlah sampel per kelas. Selain itu, dihitung metric agreget berupa akurasi data latih, akurasi data uji, serta nilai rata-rata dan simpangan baku dari hasil cross-validation. Untuk menganalisis kemampuan generalisasi model, dihitung pula train dan test gap, yaitu selisih antara akurasi pada data latih dan data uji. Visualisasi confusion matrix digunakan untuk menganalisis pola kesalahan klasifikasi antar-kelas. Selain itu, dilakukan analisis feature importance dari Random Forest untuk mengidentifikasi fitur TF-IDF yang paling berkontribusi dalam proses klasifikasi, serta analisis distribusi confidence score berupa nilai probabilitas maksimum prediksi pada data uji.

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

III. Hasil dan Pembahasan

A. Hasil Penelitian

Bagian ini menyajikan hasil pengujian dan evaluasi terhadap model klasifikasi tingkat kesulitan soal pemrograman yang dikembangkan dalam penelitian ini. Evaluasi dilakukan untuk menilai kinerja prediksi model, keseimbangan performa antar kelas, serta kemampuan generalisasi terhadap data yang belum pernah digunakan dalam proses pelatihan. Analisis hasil disusun secara bertahap, dimulai dari evaluasi performa pada masing-masing kelas tingkat kesulitan, dilanjutkan dengan analisis pola kesalahan menggunakan confusion matrix, identifikasi fitur-fitur yang paling berpengaruh melalui feature importance, serta evaluasi performa keseluruhan model menggunakan metrik akurasi dan cross-validation. Pendekatan ini bertujuan untuk memberikan gambaran yang komprehensif mengenai kekuatan dan keterbatasan model dalam konteks klasifikasi tingkat kesulitan soal berbasis teks.

Kelas	Precision	Recall	F1-Score	Support
Mudah	0.7900	0.7800	0.7845	35
Sedang	0.6800	0.6320	0.6532	37
Sulit	0.7020	0.6380	0.6712	30

Table 7. *Class Performance*

Berdasarkan Tabel 7, model menunjukkan performa yang seimbang dan cukup baik di ketiga kelas. Kelas Mudah meraih F1-score tertinggi (78,45%), diikuti oleh kelas Sulit (67,12%) dan Sedang (65,32%). Selisih performa antar kelas hanya 13,13%, mengindikasikan bahwa model tidak terlalu bias dan mampu mengenali ciri khas masing-masing tingkat kesulitan. Tingkat precision yang konsisten di atas 68% menunjukkan bahwa ketika model membuat prediksi, prediksi tersebut cenderung

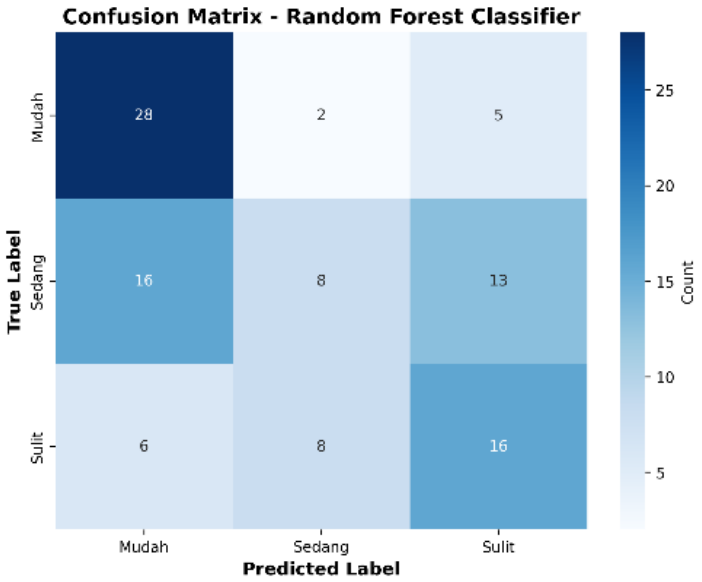


Figure 3. *Confusion Matrix*

Confusion matrix pada Gambar 3. menunjukkan bahwa sebagian besar prediksi benar berada pada elemen diagonal, dengan total 52 dari 102 data uji terklasifikasi dengan tepat. Kesalahan yang relatif kecil terjadi pada kelas Mudah dan Sedang sebanyak 2 sampel, menunjukkan adanya tumpang tindih fitur konseptual antara kedua kelas tersebut. Pola kesalahan paling dominan berasal dari kelas Sedang, khususnya 13 sampel Sedang yang diprediksi sebagai Sulit, yang mengindikasikan kecenderungan model melebihkan tingkat kesulitan pada soal dengan kompleksitas menengah. Kesalahan ekstrem, seperti Mudah dan Sulit (5 sampel), relatif jarang, menandakan bahwa model cukup mampu membedakan batas kelas ekstrem. Secara keseluruhan, terdapat sekitar 50 data yang salah diklasifikasikan dari total data uji, dengan error rate sekitar 29,02%, di mana kontribusi kesalahan terbesar berasal dari kelas Sedang, mengonfirmasi bahwa kelas ini memiliki tingkat ambiguitas tertinggi.

Indonesian Journal of Innovation Studies
Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

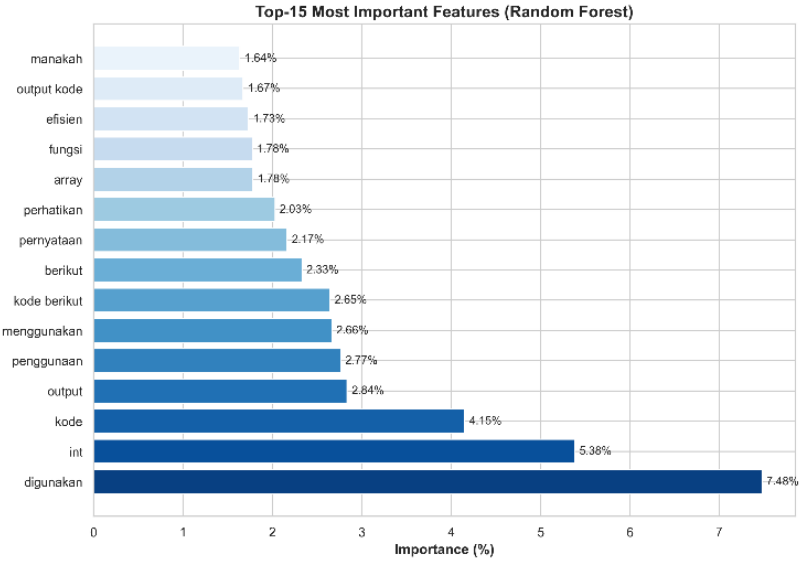


Figure 4. *Most Important Features*

Berdasarkan Gambar 4. analisis Top-15 Feature Importance menunjukkan bahwa fitur "digunakan" memiliki kontribusi tertinggi dengan nilai 7,48%, diikuti oleh "int" sebesar 5,38% dan "kode" sebesar 4,15%, yang menandakan bahwa istilah-istilah yang berkaitan langsung dengan implementasi kode dan tipe data berperan penting dalam proses klasifikasi tingkat kesulitan soal. Fitur-fitur lain seperti "output" (2,84%), "penggunaan" (2,77%), dan "menggunakan" (2,66%) juga memberikan kontribusi yang cukup konsisten, menunjukkan bahwa konteks instruksi dan penggunaan kode turut memengaruhi keputusan model. Secara keseluruhan, tiga fitur teratas menyumbang sekitar 17,01% dari total importance, sementara fitur lainnya memiliki kontribusi yang lebih kecil dan tersebar merata. Pola ini menunjukkan bahwa model memanfaatkan kombinasi beberapa fitur kunci tanpa bergantung pada satu fitur tunggal, sehingga interpretasi hasil prediksi dapat dilakukan secara lebih seimbang dan relevan dengan karakteristik soal pemrograman.

Metric	Value
Train Accuracy	82.45%
Test Accuracy	70.98%
CV Mean	69.52%
CV Std Deviation	4.12%
Train-Test Gap	11.47%

Table 8. *Performance Model*

Berdasarkan Tabel 8, evaluasi performa model menunjukkan bahwa akurasi data latih mencapai 82.45%, sedangkan akurasi data uji sebesar 70.98%, yang mengindikasikan adanya penurunan performa ketika model diterapkan pada data yang belum pernah dilihat sebelumnya. Untuk memperoleh estimasi performa yang lebih representatif, dilakukan 5-fold cross-validation yang menghasilkan rata-rata akurasi (CV Mean Score) sebesar 69,52%, sehingga nilai ini dapat digunakan sebagai gambaran kemampuan generalisasi model secara keseluruhan.

Nilai standar deviasi cross-validation sebesar 4,12% menunjukkan bahwa variasi performa antar fold berada pada tingkat sedang, yang menandakan bahwa model memberikan hasil yang relative konsisten meskipun diuji pada pembagian data yang berbeda. Perbedaan antara akurasi pelatihan dan estimasi cross-validation yang tercermin pada train dan test gap sebesar11,47% mengindikasikan bahwa karakteristik data latih dan data uji memiliki tingkat kompeksitas yang berbeda, sehingga performa model masih dopengaruhi oleh distribusi data. Secara keseluruhan, hasil ini menunjukkan bahwa cross-validation memberikan evaluasi yang lebih stabil dan realistis dibandingkan satu kali pengujian, serta menjadi dasar yang kuat untuk menilai kinerja model secara menyeluruh.

B. Pembahasan

Secara keseluruhan, Random Forest Classifier mencapai test accuracy sebesar 70,98%, yang menunjukkan peningkatan sebesar 37,65% dibandingkan baseline random guessing (33,33%) pada klasifikasi tiga kelas. Hasil ini mengindikasikan bahwa model mampu mempelajari pola yang relevan dari data teks soal. Konsistensi performa juga tercermin dari nilai rata-rata cross-validation sebesar 69.52% dengan standar deviasi 4.12%, yang menunjukkan bahwa kinerja model relatif stabil terhadap variasi pembagian data. Analisis feature importance memperlihatkan bahwa model mengandalkan sejumlah fitur

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

utama yang bermakna secara konseptual, sehingga proses pengambilan keputusan model dapat diinterpretasikan dengan cukup baik. Namun demikian, train–test gap sebesar 11.47% menunjukkan bahwa kemampuan generalisasi model masih dapat ditingkatkan, sehingga model dinilai cukup efektif sebagai alat bantu klasifikasi, tetapi belum optimal untuk penggunaan tanpa evaluasi lanjutan.

Data dari total 102 data uji, terdapat sekitar 29.02% data yang salah diklasifikasikan, dengan distribusi kesalahan yang tidak merata antar jelas. Berdasarkan evaluasi per kelas, kelas Sedang menunjukkan performa terendah dengan F1-score sebesar 65,32%, sedangkan jelas mudah memiliki F1-score tertinggisevesar 78,45%, menghasilkan selisih performa antar kelas sebesar 13,13%. Analisis confusion matrix menunjukkan bahwa kesalahan paling sering terjadi antar kelas berdekatan, khususnya antara Mudah dengan Sedang dan Sedang dengan Sulit yang mengindikasikan adanya tumpang tindih karakteristik fitur dan ambiguitas batas tingkat kesulitan. Kesalahan ekstrem antar kelas yang berjauhan relatif jarang terjadi, yang menunjukkan bahwa model masih mampu mengenali perbedaan tingkat kesulitan secara umum.

Rendahnya performa pada kelas Sedang mengindikasikan bahwa kategori kesulitan menengah merupakan level yang secara konseptual paling ambigu. Soal pada tingkat ini sering kali memiliki karakteristik linguistik dan kognitif yang tumpang tindih dengan kelas Mudah (dalam hal kompleksitas permukaan) maupun kelas Sulit (dalam hal kedalaman penalaran). Fenomena ini konsisten dengan temuan dalam literatur asesmen pendidikan, di mana kategori ordinal menegah cenderung lebih sulit dipisahkan secara tegas hanya berdasarkan fitur teks.

Berdasarkan hasil evaluasi, model ini layak digunakan sebagai pre classification tool atau decision support system, khususnya untuk membantu pengelompokan awal soal berdasarkan tingkat kesulitan. Namun, model belum disarankan untuk pengambilan keputusan berisiko tinggi tanpa intervensi manusia, mengingat tingkat kesalahan yang masih cukup signifikan, terutama pada kelas Sedang. Keterbatasan utama penelitian ini meliputi jumlah data yang relative terbatas (507 sampel) serta penggunaan fitur berbasis TF-IDF, yang belum sepenuhnya menangkap makna semantic teks. Oleh karena itu, penelitian lanjutan disarankan untuk mengeksplorasi penambahan data, penggunaan embedding semantic, serta pendekatan model yang lebih kaya guna meningkatkan keseimbangan performa antar kelas dan kemampuan generalisasi model.

Secara keseluruhan, model Random Forest berhasil menunjukkan kemampuan klasifikasi yang lebih baik daripada baseline dan mampu menangkap pola penting dalam data. Meskipun masih terdapat ruang untuk perbaikan, khususnya pada keseimbangan performa antar kelas, model ini memiliki potensi praktis sebagai alat bantu pendukung dalam klasifikasi tingkat kesulitan soal. Namun demikian, temuan utama yang paling menonjol sekaligus problematis dalam penelitian ini adalah rendahnya performa pada kelas Sedang, yang secara konsisten menunjukkan tingkat ambiguitas tertinggi baik pada evaluasi per kelas maupun analisis confusion matrix. Kondisi ini menegaskan bahwa keterbatasan model tidak semata-mata bersumber dari arsitektur algoritma, tetapi juga dari karakteristik konseptual kategori kesulitan menengah yang sulit direpresentasikan hanya melalui fitur leksikal. Oleh karena itu, arah penelitian lanjutan secara khusus perlu difokuskan pada upaya memperkaya representasi semantik dan strategi pemodelan yang mampu menangkap gradasi kognitif pada kelas Sedang, sehingga peningkatan performa di masa depan benar-benar berbasis pada permasalahan empiris yang teridentifikasi dalam penelitian ini.

IV. Kesimpulan

Penelitian ini telah berhasil mengembangkan dan mengevaluasi pipeline klasifikasi otomatis tiga tingkat kesulitan (mudah, sedang, sulit) pada soal pilihan ganda yang dihasilkan AI dengan memanfaatkan representasi teks TF-IDF dan algoritma Random Forest. Model yang dihasilkan menunjukkan kemampuan klasifikasi yang lebih baik dibandingkan baseline acak (33,33%), dengan pencapaian akurasi uji sebesar 70,98% dan konsistensi yang terukur melalui *5-fold cross-validation*. Hasil ini mengindikasikan bahwa model mampu menangkap pola linguistic tertentu yang relevan dengan persepsi kesulitan oleh pakar. Namun, analisis performa per kelas menungkapkan ketidakseimbangan yang signifikan. Kelas mudah mencapai F1-score tertinggi (78,45%), sedangkan kelas sedang mencatat performa terendah (65,32%). Temuan ini sejalan dengan tantangan umum dalam klasifikasi ordinal di bidang pendidikan, di mana kategori tengah sering kali bersifat subjektif dan kontekstual.

Dengan mempertimbangkan keterbatasan akurasi serta ketidakseimbangan performa antar kelas, model ini lebih tepat diposisikan sebagai alat bantu pra-klasifikasi atau system pendukung keputusan awal dalam proses kurasi dan penyaringan soal, bukan sebagai penentu akhir tingkat kesulitan tanpa validasi pakar manusia. Implikasi praktis ini menegaskan bahwa kontribusi utama model terletak pada efisiensi proses awal evaluasi soal, bukan pada pengambilan keputusan berisiko tinggi, Untuk penelitian selanjutnya, disarankan untuk:

1. Memperluas dataset dengan lebih banyak sampel dan variasi topik untuk meningkatkan representativitas dan generalisasi model.

2. Mengintegrasikan representasi semantik yang lebih kaya, seperti embedding berbasis transformer (misal: SBERT, IndoBERT) untuk menangkap makna kontekstual yang tidak tercover oleh TF-IDF.

3. Mengeksplorasi teknik pemodelan yang lebih kompleks, seperti ensemble methods, deep learning, atau pendekatan ordinal classification yang secara khusus dirancang untuk data berjenjang.

4. Menggabungkan fitur non-tekstual, seperti panjang soal, kompleksitas sintaksis, atau metadata kognitif, untuk memperkaya representasi soal.

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

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1880

5. Melakukan analisis kualitatif mendalam terhadap pola kesalahan klasifikasi, khususnya pada kelas sedang, sebagai upaya reflektif untuk memahami sumber ambiguitas pelabelan dan menyempurnakan kriteria evaluasi tingkat kesulitan.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa, orang tua, dan diri sendiri atas dukungan dan ketekunan selama proses penelitian. Ucapan terima kasih juga disampaikan kepada pembimbing penelitian, Bapak Utomo Pujianto, atas bimbingan dan arahan yang sangat berharga. Terima kasih kepada Ibu Putrinda Inayatul Maula dan Bapak Muzayyin Amrulloh atas kontribusi dalam anotasi data dan validasi ground truth. Tidak lupa, penulis berterima kasih kepada semua teman yang telah memberikan dukungan, diskusi, dan bantuan selama pelaksanaan penelitian ini

References

1. N. Scaria, S. D. Chenna, and D. Subramani, “Automated Educational Question Generation at Different Bloom’s Skill Levels using Large Language Models: Strategies and Evaluation,” vol. 14830, 2024, pp. 165–179. doi: 10.1007/978-3-031-64299-9_12.
2. I. L. Molina, V. Švábenský, T. Minematsu, L. Chen, F. Okubo, and A. Shimada, “Comparison of Large Language Models for Generating Contextually Relevant Questions,” vol. 15160, 2024, pp. 137–143. doi: 10.1007/978-3-031-72312-4_18.
3. R. Fulari and J. Rusert, “Utilizing Machine Learning to Predict Question Difficulty and Response Time for Enhanced Test Construction”.
4. A. D. M. Putri, N. Sulistianingsih, and R. Rismayati, “Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain,” JTIM J. Teknol. Inf. Dan Multimed., vol. 7, no. 4, pp. 675–688, Oct. 2025, doi: 10.35746/jtim.v7i4.756.
5. K. Madatov, S. Sattarova, and J. Vičič, “TF-IDF-Based Classification of Uzbek Educational Texts,” Appl. Sci., vol. 15, no. 19, p. 10808, Oct. 2025, doi: 10.3390/app151910808.
6. Y. Dai, F. Wang, and J. Luo, “Optimal Opacity-Enforcing Supervisory Control of Discrete Event Systems on Choosing Cost,” Appl. Sci., vol. 14, no. 6, p. 2532, Mar. 2024, doi: 10.3390/app14062532.
7. P. S. Siregar, “Multiple Choice Question Difficulty Level Classification with Multi Class Confusion Matrix in the Online Question Bank of Education Gallery,” J. Appl. Data Sci., vol. 4, no. 4, pp. 392–406, Dec. 2023, doi: 10.47738/jads.v4i4.132.
8. M. R. Syaputra, M. Arifin, and D. L. Fithri, “Klasifikasi Sentimen Ulasan E-Wallet menggunakan TF-IDF dan Random Forest dengan Penyeimbangan Data SMOTE”.
9. R. Gupta, R. Aksitov, S. Phatale, S. Chaudhary, H. Lee, and A. Rastogi, “Conversational Recommendation as Retrieval: A Simple, Strong Baseline,” May 23, 2023, arXiv: arXiv:2305.13725. doi: 10.48550/arXiv.2305.13725.
10. M. L. McHugh, “Interrater reliability: the kappa statistic,” Biochem. Medica, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.
11. G. I. Kim, S. Kim, and B. Jang, “Classification of mathematical test questions using machine learning on datasets of learning management system questions,” PLOS ONE, vol. 18, no. 10, p. e0286989, Oct. 2023, doi: 10.1371/journal.pone.0286989.
12. Jubeile Mark Baladjay, Nisce Riva, Ladine Ashley Santos, Dan Michael Cortez, Criselle Centeno, and Ariel Antwaun Rolando Sison, “Performance evaluation of random forest algorithm for automating classification of mathematics question items,” World J. Adv. Res. Rev., vol. 18, no. 2, pp. 034–043, May 2023, doi: 10.30574/wjarr.2023.18.2.0762.
13. C. Isley et al., “Assessing the Quality of AI-Generated Exams: A Large-Scale Field Study,” Aug. 09, 2025, arXiv: arXiv:2508.08314. doi: 10.48550/arXiv.2508.08314.
14. A. Yaacoub, J. Da-Rugna, and Z. Assaghir, “Assessing AI-Generated Questions’ Alignment with Cognitive Frameworks in Educational Assessment”.
15. S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma, “Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches,” Int. J. Artif. Intell. Educ., vol. 34, no. 3, pp. 862–914, Sept. 2024, doi: 10.1007/s40593-023-00362-1.
16. L. Zotos, I. P. de Jong, M. Valdenegro-Toro, A. I. Sburlea, M. Nissim, and H. van Rijn, “NLP Methods May Actually Be Better Than Professors at Estimating Question Difficulty,” 2025, arXiv. doi: 10.48550/ARXIV.2508.03294.
17. R. S. Perdana and P. P. Adikara, “Multi-task Learning for Named Entity Recognition and Intent Classification in Natural Language Understanding Applications,” J. Inf. Syst. Eng. Bus. Intell., vol. 11, no. 1, pp. 1–16, Mar. 2025, doi: 10.20473/jisebi.11.1.1-16.
18. M. Li, Q. Gao, and T. Yu, “Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters,” BMC Cancer, vol. 23, no. 1, p. 799, Aug. 2023, doi: 10.1186/s12885-023-11325-z.
19. M. Méndez, M. G. Merayo, and M. Núñez, “Design of hybrid machine learning and TF-IDF models to discard irrelevant reviews on public transport stations,” J. Inf. Telecommun., vol. 9, no. 4, pp. 481–504, Oct. 2025, doi: 10.1080/24751839.2025.2472503.
20. P. Guleria, J. Frnda, and P. N. Srinivasu, “NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis,” Array, vol. 27, p. 100467, Sept. 2025, doi: 10.1016/j.array.2025.100467.