Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

Error Checker Berbasis Neural Network pada Pembentukan Soal Cerita Matematika Dasar

Ferri Nurdiansah[1], Agung Prasetya[2], Taufiq Agung Cahyono[3]
[1,2,3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Bhinneka PGRI, Tulungagung, Indonesia
Email: [1]ferrinurdiansah@gmail.com, [2]agung@ubhi.ac.id, [3]taufiqagungcahyono@gmail.com

Abstrak

Soal cerita matematika (*Math Word Problem*/MWP) berperan penting dalam melatih kemampuan penalaran logis siswa Sekolah Dasar. Perkembangan *Large Language Model* (LLM) telah memungkinkan pembuatan soal secara otomatis, namun soal yang dihasilkan kerap mengandung berbagai jenis kesalahan. Penelitian ini bertujuan: (1) mengembangkan dataset MWP berbahasa Indonesia beranotasi 11 kategori kesalahan secara multi-label; (2) merancang model *multi-label classification* berbasis IndoBERT dan *Feed Forward Neural Network* (FFNN) untuk mendeteksi kesalahan secara simultan; serta (3) mengevaluasi kinerja model menggunakan metrik Micro F1-Score, Macro F1-Score, dan Hamming Loss. Dataset dikembangkan melalui *error-inducing prompting*, menghasilkan 2.517 soal cerita yang dianotasi secara manual oleh dua anotator dengan rasio pembagian 80:20 menggunakan *stratified random sampling*. Model dilatih selama 40 epoch dengan *Binary Cross Entropy loss* dan optimizer AdamW. Model optimal diperoleh pada epoch ke-13 dengan *validation loss* 0,2616. Evaluasi pada *testing set* menghasilkan Micro F1-Score sebesar 0,7657, Macro F1-Score sebesar 0,7631, dan Hamming Loss sebesar 0,1387. *Topic Unsafety* mencapai F1-Score sempurna (1,000) dan *Misspellings* mendekati sempurna (0,986), sementara *Grade Mismatch* menjadi kategori paling sulit dideteksi (F1 = 0,591). Penelitian ini menyimpulkan bahwa pendekatan *multi-label classification* berbasis IndoBERT efektif untuk deteksi error pada MWP berbahasa Indonesia.
Kata Kunci: Error Checker; IndoBERT; Multi-Label Classification; Neural Network; Soal Cerita Matematika

Abstract

Math Word Problems (MWPs) play a crucial role in developing the logical reasoning skills of elementary school students. Recent advances in Large Language Models (LLMs) have enabled automated generation of such problems; however, LLM-generated MWPs frequently contain various types of errors. This study aims to: (1) develop an Indonesian-language MWP dataset annotated with 11 error categories in a multi-label scheme; (2) design a multi-label classification model based on IndoBERT and a Feed Forward Neural Network (FFNN) for simultaneous error detection; and (3) evaluate model performance using Micro F1-Score, Macro F1-Score, and Hamming Loss metrics. The dataset was developed through error-inducing prompting, yielding 2,517 Indonesian-language MWPs manually annotated by two annotators, split 80:20 using stratified random sampling. The model was trained for 40 epochs using Binary Cross Entropy loss and AdamW optimizer. The optimal model was obtained at epoch 13 with a validation loss of 0.2616. Evaluation on the testing set yielded a Micro F1-Score of 0.7657, Macro F1-Score of 0.7631, and Hamming Loss of 0.1387. Topic Unsafety achieved a perfect F1-Score of 1.000 and Misspellings near-perfect (0.986), while Grade Mismatch was the most challenging category (F1 = 0.591). This study concludes that the IndoBERT-based multi-label classification approach is effective for error detection in Indonesian-language MWPs.
Keywords: Error Checker; IndoBERT; Math Word Problem; Multi-Label Classification; Neural Network

***Corresponding Author:**
Ferri Nurdiansah, Universitas Bhinneka PGRI, Indonesia
Email: ferrinurdiansah@gmail.com

1. PENDAHULUAN

Matematika pada tingkat Sekolah Dasar tidak semata-mata berkaitan dengan angka, melainkan menekankan pada pemahaman konteks. Soal cerita atau *Math Word Problem* (MWP) memegang peranan

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

penting dalam melatih kemampuan penalaran logis siswa serta mendorong mereka mentransformasikan permasalahan kehidupan sehari-hari ke dalam bentuk bahasa matematis [1]. Secara umum, MWP didefinisikan sebagai representasi masalah matematika dalam bentuk narasi yang memerlukan pemahaman bahasa tingkat tinggi untuk mengekstrak informasi numerik dan menentukan operasi matematika yang tepat [6]. Oleh karena itu, penyusunan MWP yang baik memerlukan perhatian ketat pada kejelasan narasi.

Perkembangan *Large Language Model* (LLM) telah mendominasi berbagai bidang, termasuk pembuatan soal cerita secara otomatis. Beberapa penelitian telah menunjukkan potensi LLM dalam konteks ini: Ariyarathne et al. [2] memastikan soal buatan LLM dapat diselesaikan dengan *Solvability Checker*; Zhou dan Huang [3] memelopori pembangkitan MWP berbasis topik menggunakan model neural; Kang et al. [4] mengembangkan *framework* berbasis *template* dan LLM untuk menghasilkan tabular MWP yang beragam dan akurat; model MATHWELL (Christ et al. [5]) melibatkan guru sungguhan untuk menilai kualitas pedagogis soal; He-Yueya et al. [7] mengombinasikan LLM dengan *symbolic solver* untuk meningkatkan akurasi penyelesaian; dan Shridhar et al. [8] mengembangkan pembangkitan subpertanyaan Socratic untuk memandu pemahaman siswa.

Meskipun demikian, LLM sering menghasilkan soal yang secara perhitungan dapat diselesaikan namun masih mengandung permasalahan pada aspek teks. Kim et al. [9] memperkenalkan DMath, *dataset* MWP beragam yang membuktikan GPT-4 pun hanya mencapai akurasi ~75%, menunjukkan kualitas soal yang dihasilkan LLM masih sangat bervariasi; Xie et al. [10] membuktikan bahwa soal adversarial dapat mengecoh model penyelesai MWP. Keterbatasan LLM dalam menghasilkan teks yang valid secara linguistik dan pedagogis ini memicu munculnya berbagai kesalahan pada draf soal. Kesalahan tersebut umumnya mencakup ke dalam 11 kategori kesalahan pada MWP, yaitu: (1) *Co-Reference Issues*, (2) *Trivial Problems*, (3) *Grammatical Errors*, (4) *Misspellings*, (5) *Incomplete Sentences*, (6) *Unsolvable Problems*, (7) *Unrealistic Problems*, (8) *Unit Issues*, (9) *Topic Unsafety*, (10) *Grade Mismatch*, dan (11) *Not a Word Problem*.

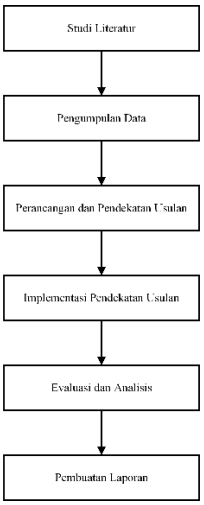
Penelitian ini mengusulkan pendekatan *multi-label classification* untuk mendeteksi kesalahan pada MWP berbahasa Indonesia secara simultan. Pendekatan *multi-label* dipilih karena satu buah soal yang dihasilkan LLM sangat mungkin memiliki lebih dari satu jenis kesalahan sekaligus secara bersamaan, sejalan dengan keberhasilan metode ini pada berbagai penelitian NLP terdahulu [13-17]. Sistem ini menggunakan IndoBERT sebagai *encoder* dan *Feed Forward Neural Network* (FFNN) sebagai lapisan klasifikasi. IndoBERT digunakan sebagai komponen penghela utama karena merupakan model berbasis *transformer* [11, 12] yang telah dilatih khusus pada korpus Bahasa Indonesia berskala besar, sehingga mampu menghasilkan representasi vektor kontekstual yang akurat. Efektivitas IndoBERT pada tugas klasifikasi jamak berbahasa Indonesia juga telah dibuktikan secara empiris pada domain lain, seperti peningkatan F1-Score hingga 6,17% atas baseline pada klasifikasi ulasan pelanggan [13] dan pencapaian F1-Score 0,9032 pada deteksi komentar toksik [14], sehingga menguatkan alasan pemilihannya sebagai encoder utama dalam penelitian ini. Fokus pada Bahasa Indonesia ini merupakan kontribusi signifikan mengingat dominasi penelitian evaluasi LLM yang selama ini masih berbasis Bahasa Inggris. Berdasarkan permasalahan tersebut, tujuan spesifik dari penelitian ini adalah merancang dan mengimplementasikan arsitektur model *error checker* otomatis berbasis integrasi IndoBERT-FFNN, serta mengevaluasi kinerjanya secara komprehensif dalam mendeteksi 11 kategori kesalahan pada *dataset* berlabel jamak berbahasa Indonesia.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan secara sistematis melalui beberapa tahapan utama untuk menjamin ketercapaian tujuan penelitian. Alur pengerjaan penelitian secara menyeluruh ditunjukkan melalui diagram alir pada Gambar 1.

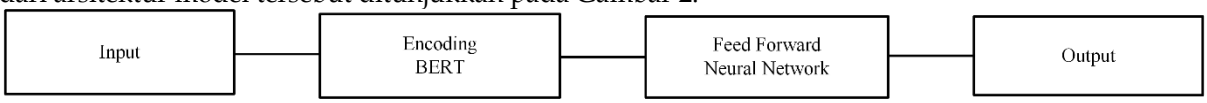


Gambar 1. Diagram Alur Tahapan Penelitian

Proses diawali dengan **studi literatur** untuk mengkaji berbagai penelitian terdahulu yang relevan dengan topik penyusunan soal cerita matematika serta pendeteksian kesalahan. Tahap berikutnya adalah **pengumpulan data** berupa soal cerita matematika (*Math Word Problem*) yang bersumber dari hasil keluaran *Large Language Model* (LLM). Setelah data terkumpul, dilakukan **perancangan pendekatan** dengan memanfaatkan IndoBERT sebagai *encoder* dan metode *multi-label classification* untuk mengklasifikasikan berbagai jenis kesalahan yang mungkin muncul. Rancangan tersebut kemudian direalisasikan pada tahap **implementasi pendekatan** ke dalam bentuk program yang mampu memproses soal cerita dan melakukan prediksi untuk mengidentifikasi kesalahan pada setiap kategori yang telah ditentukan. Selanjutnya, dilakukan tahap **evaluasi dan analisis** terhadap hasil pengujian model dengan menggunakan beberapa metrik kinerja, antara lain *precision, recall, F1-score,* dan *hamming loss*. Rangkaian kegiatan ini diakhiri dengan **pembuatan laporan** penelitian sebagai bentuk dokumentasi yang sistematis dan komprehensif dari seluruh pengerjaan yang telah dilakukan.

2.2 Arsitektur Model

Arsitektur model yang diusulkan dalam penelitian ini dirancang untuk menangani klasifikasi jamak (*multi-label classification*) pada teks soal cerita matematika secara simultan. Secara umum, diagram blok dari arsitektur model tersebut ditunjukkan pada Gambar 2.



Gambar 2. Arsitektur Multi Label untuk Deteksi Error

Berdasarkan Gambar 2, alur pemrosesan model terbagi menjadi beberapa tahapan terintegrasi. Proses diawali dengan pemberian masukan (*input*) berupa teks soal cerita matematika yang dihasilkan oleh *Large Language Model* (LLM). Langkah kedua adalah proses *encoding* teks menggunakan IndoBERT, di mana teks soal cerita diubah menjadi representasi vektor kontekstual untuk menangkap makna linguistik secara mendalam.

Vektor hasil keluaran IndoBERT kemudian dimasukkan ke dalam lapisan *dense* pada *Feed Forward Neural Network* (FFNN) untuk dipetakan ke dalam ruang label, di mana setiap dimensi merepresentasikan satu kategori kesalahan. Selanjutnya, keluaran dari FFNN diproses menggunakan fungsi aktivasi sigmoid untuk menghasilkan nilai probabilitas pada rentang 0 hingga 1 yang

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

menunjukkan tingkat keyakinan model. Tahap terakhir adalah penentuan *output* label kesalahan melalui mekanisme *thresholding* dengan nilai ambang batas 0,5. Pendekatan ini memungkinkan model menghasilkan vektor *multilabel* untuk mendeteksi berbagai jenis kesalahan secara bersamaan dalam satu soal cerita.

2.3 Konfigurasi Pelatihan

Model diimplementasikan menggunakan bahasa pemrograman Python dengan *framework* PyTorch serta pustaka Hugging Face Transformers. Pelatihan dilakukan selama 40 *epoch* dengan fungsi *loss Binary Cross Entropy* (BCE) dan *optimizer* AdamW. Model terbaik disimpan berdasarkan nilai *validation loss* terendah. Konfigurasi lengkap mengenai parameter-parameter yang digunakan dalam proses pelatihan model ini dicantumkan secara rinci pada Tabel 1.

Tabel 1. Konfigurasi Pelatihan Model

Parameter	Nilai
Encoder	IndoBERT-base
Jumlah Epoch	40
Fungsi Loss	Binary Cross Entropy (BCE)
Optimizer	AdamW
Threshold	0,5
Jumlah Label	11
Strategi Simpan Model	*Best Validation Loss*
Framework	PyTorch + Hugging Face Transformers

Berdasarkan Tabel 1, konfigurasi parameter pelatihan ditentukan untuk mengoptimalkan kinerja model IndoBERT-base dalam mengenali karakteristik teks soal cerita matematika. Penggunaan 11 label luaran disesuaikan dengan jumlah kategori kesalahan yang didefinisikan dalam penelitian ini. Fungsi *loss Binary Cross Entropy* (BCE) dipilih karena karakteristiknya yang optimal untuk skenario klasifikasi jamak (*multi-label classification*), di mana setiap kategori kesalahan dihitung probabilitasnya secara independen menggunakan fungsi aktivasi sigmoid dengan batas ambang (*threshold*) sebesar 0,5. Guna meminimalkan risiko terjadinya *overfitting* selama 40 *epoch* masa pelatihan, digunakan *optimizer* AdamW untuk memperbarui bobot model secara efisien, serta diterapkan strategi penyimpanan model terbaik yang dievaluasi berdasarkan pencapaian nilai *validation loss* paling minimum.

3. HASIL DAN PEMBAHASAN

3.1. Deskripsi Dataset

Dataset yang berhasil dikumpulkan terdiri dari 2.517 soal cerita berbahasa Indonesia. Sebanyak 1.330 soal (52,8%) merupakan soal dengan *multi-label*, yang mengonfirmasi keputusan penggunaan pendekatan *multi-label classification*. Distribusi label relatif seimbang di antara semua kategori, dengan *Misspellings* memiliki jumlah tertinggi (818 kemunculan) dan *Grade Mismatch* terendah (597 kemunculan).

Dataset dikembangkan secara semi-sintetis melalui proses *error-inducing prompting* menggunakan LLM ringan Gemma3:1b yang dijalankan secara lokal. Setiap permintaan menghasilkan 10 soal cerita dengan variasi nama tokoh, konteks, angka, dan operasi matematika. Prompt generasi secara eksplisit mengarahkan model untuk menyisipkan 2 hingga 4 kesalahan alami pada setiap soal, tanpa menargetkan kategori kesalahan tertentu secara spesifik, sehingga jenis kesalahan yang muncul bersifat alami mengikuti kecenderungan LLM itu sendiri. Soal-soal yang dihasilkan selanjutnya melalui proses anotasi manual oleh dua anotator berlatar belakang informatika dan matematika, yang mengklasifikasikan setiap soal ke dalam satu atau lebih dari 11 kategori kesalahan yang telah didefinisikan.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

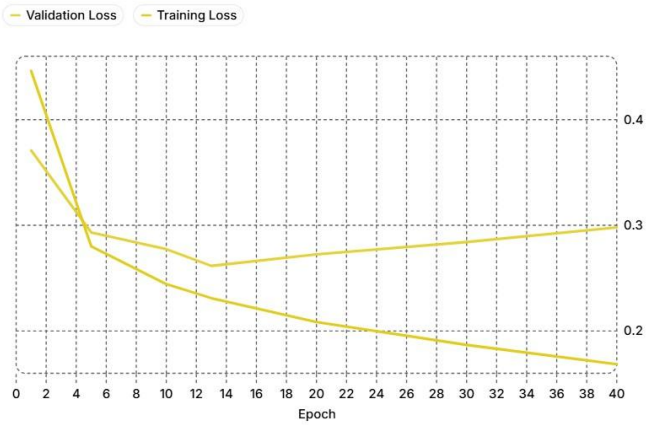
Tabel 2. Distribusi Label pada Dataset

No	Label Error	Jumlah	Persentase (%)
1	*Misspellings*	818	9,87
2	*Trivial Problem*	811	9,79
3	*Grammatical Error*	810	9,77
4	*Co-Reference Issues*	793	9,57
5	*Unrealistic Problem*	787	9,50
6	*Unit Issue*	776	9,36
7	*Unsolvable Problem*	768	9,27
8	*Topic Unsafety*	766	9,24
9	*Incomplete Sentence*	742	8,95
10	*Not a Word Problem*	618	7,46
11	*Grade Mismatch*	597	7,20
	Total	8.287	100,00

Berdasarkan Tabel 2, distribusi label pada dataset relatif seimbang di antara semua kategori. Label *Misspellings* memiliki jumlah tertinggi dengan 818 kemunculan, sedangkan *Grade Mismatch* menjadi yang terendah dengan 597 kemunculan. Ketimpangan antar label masih berada dalam batas yang wajar dan tidak berpotensi menyebabkan bias signifikan selama proses pelatihan. Kondisi distribusi yang seimbang ini menjadi salah satu kekuatan dataset yang dikembangkan dalam penelitian ini.

3.2. Hasil Pelatihan Model

Model dilatih secara bertahap selama 40 *epoch*. Tren perubahan nilai *loss* selama proses pelatihan dan validasi untuk setiap *epoch* disajikan secara visual pada Gambar 3.



Gambar 3. Hasil Pelatihan Model

Berdasarkan Gambar 3, nilai *training loss* menunjukkan penurunan yang konsisten dari 0,4462 pada *epoch* pertama hingga mencapai 0,1686 pada *epoch* ke-40. Sementara itu, nilai *validation loss* mencapai titik terendah pada *epoch* ke-13 sebesar 0,2616. Strategi *best model saving* berdasarkan *validation loss* terendah terbukti efektif untuk mengunci bobot model pada titik optimal tersebut sebelum performanya mengalami penurunan. Nilai *Micro F1-Score* dan *Macro F1-Score* turut mencapai puncak pada *epoch* ke-13, masing-masing sebesar 0,7783 dan 0,7735, dengan *Hamming Loss* terendah sebesar 0,1333. Melalui kondisi tersebut, *epoch* ke-13 ditetapkan sebagai *epoch* optimal.

3.3. Evaluasi pada Testing Set

Evaluasi akhir dilakukan pada *testing set* menggunakan model terbaik yang tersimpan pada epoch ke-13. Hasil evaluasi secara keseluruhan ditampilkan pada Tabel 3

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

Tabel 3. Hasil Evaluasi Model pada Testing Set

Metrik	Nilai
Micro Precision	0,7393
Micro Recall	0,7940
Micro F1-Score	0,7657
Macro Precision	0,7411
Macro Recall	0,7887
Macro F1-Score	0,7631
Hamming Loss	0,1387
Samples Avg F1-Score	0,8176

Nilai Micro F1-Score sebesar 0,7657 menunjukkan kinerja yang baik secara keseluruhan. Nilai Recall (0,7940) yang lebih tinggi daripada Precision (0,7393) mengindikasikan model lebih cenderung mendeteksi seluruh *error* yang ada, meskipun ada beberapa *false positive*. Kedekatan Micro dan Macro F1-Score (selisih 0,0026) mengindikasikan tidak ada bias signifikan terhadap label tertentu. *Hamming Loss* sebesar 0,1387 menunjukkan rata-rata hanya 13,87% label diprediksi secara keliru.

Sebagai pembanding, penelitian *multi-label classification* berbasis IndoBERT pada domain lain melaporkan performa yang lebih tinggi, seperti F1-Score 0,9032 pada deteksi komentar toksik [14] dan peningkatan F1-Score hingga 6,17% atas baseline pada klasifikasi ulasan pelanggan [13]. Nilai Micro F1-Score 0,7657 pada penelitian ini relatif lebih rendah, yang dapat dijelaskan oleh karakteristik tugas yang berbeda: deteksi error MWP menuntut pemahaman semantik dan pedagogis yang lebih kompleks dibanding klasifikasi topik atau sentimen, karena sebagian kategori kesalahan (seperti *Grade Mismatch* dan *Unrealistic Problem*) bersifat subjektif dan memerlukan penalaran kontekstual mendalam. Sejauh penelusuran penulis, belum terdapat penelitian pembanding yang secara langsung mengevaluasi deteksi error multi-label pada MWP berbahasa Indonesia, sehingga hasil penelitian ini dapat menjadi titik acuan awal (baseline) bagi penelitian sejenis di masa mendatang.

3.4. Evaluasi Per Kategori Label

Evaluasi per kategori label dilakukan untuk mengidentifikasi kategori error yang paling mudah dan paling sulit dideteksi oleh model. Hasil evaluasi per label ditampilkan pada Tabel 4

Tabel 4. Hasil Evaluasi Model per Kategori Label

Label	Precision	Recall	F1-Score	Support
Co-Reference Issues	0,6750	0,7347	0,7036	147
Grade Mismatch	0,5645	0,6195	0,5907	113
Grammatical Error	0,5894	0,7439	0,6577	164
Incomplete Sentence	0,6763	0,6395	0,6573	147
Misspellings	0,9718	1,0000	0,9857	172
Not a Word Problem	0,8029	0,8333	0,8178	132
Topic Unsafety	1,0000	1,0000	1,0000	137
Trivial Problem	0,8125	0,8125	0,8125	176
Unit Issue	0,7879	0,8497	0,8176	153
Unrealistic Problem	0,5988	0,6944	0,6431	144
Unsolvable Problem	0,6726	0,7483	0,7085	151
Macro Avg	0,7411	0,7887	0,7631	1.636

Berdasarkan data pada Tabel 4, dilakukan analisis performa model untuk setiap kategori kesalahan. *Topic Unsafety* mencapai *F1-Score* sempurna (1,000) karena memiliki pola linguistik yang sangat khas dan berbeda. *Misspellings* mendekati sempurna (F1 = 0,986) dengan *Recall* sempurna (1,000) karena fitur

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.316
Page 487-494

leksikalnya yang jelas dan unik. *Not a Word Problem* (F1 = 0,8178) dan *Unit Issue* (F1 = 0,8176) juga menunjukkan performa sangat baik.

Grade Mismatch merupakan kategori paling sulit dengan F1-Score 0,5907, karena penilaian kesesuaian jenjang memerlukan pemahaman semantik mendalam yang bersifat subjektif. *Unrealistic Problem* (F1 = 0,6431) sulit dinilai karena membutuhkan pengetahuan dunia nyata dan konteks budaya. *Grammatical Error* (F1 = 0,6577) mengalami kesulitan akibat kompleksitas variasi tata bahasa Indonesia.

3.5. Fenomena Overfitting

Berdasarkan tren kurva yang disajikan pada Gambar 3, analisis kurva pelatihan menunjukkan indikasi *overfitting* yang dimulai setelah *epoch* ke-13. Fenomena ini ditandai secara jelas oleh pergerakan nilai *training loss* yang terus menurun secara konstan, sedangkan nilai *validation loss* mulai berbalik menunjukkan tren peningkatan yang menjauh dari garis *training*. Meskipun terjadi gejala tersebut, penerapan strategi *best model saving* berhasil mengantisipasi masalah dengan mengamankan parameter model pada kondisi paling optimal di *epoch* ke-13. Untuk pengembangan ke depannya, implementasi teknik regularisasi seperti peningkatan nilai *dropout*, penerapan *weight decay*, atau penggunaan mekanisme *early stopping* yang lebih agresif dapat dipertimbangkan guna meminimalkan celah generalisasi antara data pelatihan dan data validasi.

4. KESIMPULAN

Penelitian ini berhasil: (1) mengembangkan dataset 2.517 MWP berbahasa Indonesia beranotasi 11 kategori kesalahan secara *multi-label* melalui *error-inducing prompting*; (2) membangun model *multi-label classification* berbasis IndoBERT yang mencapai Micro F1-Score 0,7657, Macro F1-Score 0,7631, dan *Hamming Loss* 0,1387; serta (3) mengidentifikasi pola performa per kategori, di mana kesalahan leksikal-struktural lebih mudah dideteksi dibanding kesalahan kontekstual-pedagogis.

Secara khusus, penelitian ini menyediakan *benchmark* awal untuk deteksi error multi-label pada soal cerita matematika berbahasa Indonesia berbasis IndoBERT, sehingga mengisi kekosongan penelitian evaluasi kualitas MWP otomatis yang selama ini didominasi oleh Bahasa Inggris.

Penelitian ini membuktikan bahwa pendekatan *multi-label classification* berbasis IndoBERT merupakan solusi yang layak untuk membangun sistem *error checker* otomatis pada MWP berbahasa Indonesia yang dihasilkan LLM, dengan potensi menjadi komponen *quality assurance* dalam ekosistem pembuatan soal berbasis AI untuk pendidikan dasar.

Untuk penelitian selanjutnya disarankan: memperluas dataset dengan soal buatan guru; meningkatkan proses anotasi dengan lebih banyak anotator dan pengukuran *Cohen's Kappa*; mengoptimalkan *threshold* per label; menerapkan teknik regularisasi lebih kuat; mengeksplorasi model IndoBERT-large atau *ensemble learning*; dan mengembangkan sistem *end-to-end* yang juga mampu memberikan rekomendasi perbaikan.

REFERENCES

[1] A. Prasetya, "Identifying arithmetic operations in math word problems based on recursive neural network and support vector machine," J. Online Inform., vol. 8, no. 2, 2024, doi: 10.29100/joeict.v8i2.7421.
[2] S. Ariyarathne, D. Athapaththu, S. Ekanayake, and S. Ranathunga, "Elementary math word problem generation using large language models," arXiv:2506.05950, 2025, doi: 10.48550/arXiv.2506.05950.
[3] Q. Zhou and D. Huang, "Towards generating math word problems from equations and topics," in Proc. 12th Int. Conf. Nat. Lang. Gener. (INLG 2019), Tokyo, Japan, 2019, pp. 494-503, doi: 10.18653/v1/W19-8661.
[4] X. Kang, Z. Wang, X. Jin, W. Wang, K. Huang, and Q. Wang, "Template-driven LLM-paraphrased framework for tabular math word problem generation," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 23, 2025, pp. 24303-24311, doi: 10.1609/aaai.v39i23.34607.
[5] T. Christ, P. Bhatt, S. Bhat, and A. Lan, "MATHWELL: Generating educational math word problems at scale," in Findings of EMNLP 2024, Miami, FL, USA, 2024, pp. 1598-1613, doi: 10.48550/arXiv.2402.15861.
[6] S. Mandal and S. K. Naskar, "Classifying and solving arithmetic math word problems: An intelligent math solver," IEEE Trans. Learn. Technol., vol. 14, no. 1, pp. 28-41, Feb. 2021, doi: 10.1109/TLT.2021.3057805.

[7] J. He-Yueya, G. Poesia, R. E. Wang, and N. D. Goodman, "Solving math word problems by combining language models with symbolic solvers," arXiv:2304.09102, 2023, doi: 10.48550/arXiv.2304.09102.
[8] K. Shridhar, J. Macina, M. El-Assady, T. Sinha, M. Kapur, and M. Sachan, "Automatic generation of socratic subquestions for teaching math word problems," in Proc. EMNLP 2022, Abu Dhabi, UAE, 2022, pp. 4136-4149, doi: 10.18653/v1/2022.emnlp-main.277.
[9] J. Kim, Y. Kim, I. Baek, J. Bak, and J. Lee, "It ain't over: A multi-aspect diverse math word problem dataset," in Proc. 2023 Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Singapore, 2023, pp. 14984-15011, doi: 10.18653/v1/2023.emnlp-main.927.
[10] Z. Xie, K. Zhou, S. Shi, and S. Shi, "Adversarial math word problem generation," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 17, 2024, pp. 19238-19246, doi: 10.1609/aaai.v38i17.29900.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT 2019, Minneapolis, MN, USA, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.
[12] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in Proc. 28th COLING, Barcelona, Spain, 2020, pp. 757-770, doi: 10.18653/v1/2020.coling-main.66.
[13] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," Int. J. Electr. Comput. Eng., vol. 13, no. 5, pp. 5641-5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
[14] G. Z. Nabiilah, I. Nur, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," Int. J. Electr. Comput. Eng., vol. 14, no. 1, pp. 1071-1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
[15] N. C. Mei, S. Tiun, and G. Sastria, "Multi-label aspect-sentiment classification on Indonesian cosmetic product reviews with IndoBERT model," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 11, 2024, doi: 10.14569/IJACSA.2024.0151168.
[16] Y. Sagama and A. Alamsyah, "Multi-label classification of Indonesian online toxicity using BERT and RoBERTa," in Proc. IAICT 2023, Bandung, Indonesia, 2023, pp. 264-269, doi: 10.1109/IAICT59002.2023.10205892.
[17] Y. K. Sari, J. Iskandar, and A. Prasetya, "Penerapan metode Naive Bayes dalam analisis sentimen terhadap cyberbullying," JUSTER: J. Sains dan Terapan, vol. 5, no. 1, pp. 64-73, 2026, doi: 10.57218/juster.v5i1.2556.