HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 2, Desember 2026, p392-399
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no2

RANCANG BANGUN *CHATBOT* PEMBELAJARAN BAHASA INGGRIS MENGGUNAKAN PENDEKATAN *HYBRID ARTIFICIAL INTELLIGENCE* UNTUK MENGATASI *FOSSILIZED ERROR* PADA PELAJAR INDONESIA

R. Syailindra Alfarraiel[1], Muhammad Rafly Wahyudi[2], Yohanes Olvin Jun Sole[3], Shindi Shella May Wara[4], Amri Muhaimin[5]

Program Studi Sains Data, UPN "Veteran" Jawa Timur
Jl. Raya Rungkut Madya, Gunung Anyar, Surabaya - Jawa Timur, Indonesia
Email: [1]24083010030@studentupnjatim.ac.id, [2]24083010078@studentupnjatim.ac.id
[3]24083010029@studentupnjatim.ac.id, [4]shindi.shella.fasilkom@upnjatim.ac.id,
[5]amri.muhaimin.stat@upnjatim.ac.id

ABSTRAK

Penguasaan bahasa Inggris masyarakat Indonesia masih tergolong rendah, sebagaimana ditunjukkan oleh laporan *EF English Proficiency Index* (EF EPI) 2023 yang menempatkan Indonesia pada peringkat ke-79 dari 113 negara dengan kategori *low proficiency*. Salah satu akar permasalahannya adalah fenomena *fossilized error*, yaitu kesalahan gramatikal yang menetap dalam jangka panjang akibat pengaruh bahasa pertama (L1). Sebagai upaya mengatasi permasalahan tersebut, dikembangkan sistem *chatbot* berbasis Hybrid AI yang mengintegrasikan sistem berbasis aturan (*rule-based*), *Natural Language Processing* (NLP), dan basis pengetahuan (*knowledge base*) untuk mendeteksi serta mengoreksi kesalahan gramatikal bahasa Inggris pada penutur Indonesia. Sistem dikembangkan menggunakan Python dengan *library SentenceTransformers,* FAISS, dan algoritma Naive Bayes untuk klasifikasi kesalahan gramatikal ke dalam 36 kategori. Evaluasi dilakukan menggunakan *Black Box* Testing dengan metrik akurasi, presisi, recall, dan F1-score. Hasil evaluasi menunjukkan Naive Bayes mencapai akurasi klasifikasi sebesar 32,18% pada 36 kelas *error* yang sangat granular, dengan nilai tersebut 11,6 kali lebih baik dari baseline acak (2,78%). Kelas terbaik mencapai F1-*score* 0,82 pada kategori *Negation Errors, Quantifier Errors,* dan *Mixed Metaphors/Idioms.* Rata-rata skor kemiripan semantik sistem (Avg *Similarity Score*) mencapai 42,17%, menunjukkan kemampuan sistem dalam memberikan respons yang relevan secara kontekstual. Fitur *Text to Speech* dan *Upload File* terbukti meningkatkan pengalaman belajar pengguna. Kontribusi yang dihasilkan meliputi prototipe *chatbot* edukatif berbasis web yang dapat diakses secara daring beserta *dataset fossilized error* khas penutur Indonesia.

Kata kunci: *chatbot, fossilized error, Hybrid* AI, NLP, pembelajaran bahasa Inggris

ABSTRACT

English proficiency among Indonesians remains considerably low, as indicated by the EF English Proficiency Index (EF EPI) 2023 report, which ranked Indonesia 79th out of 113 countries in the low proficiency category . One of the root causes is the phenomenon of fossilized errors, that is, grammatical mistakes that persist long-term due to the influence of the learner's first language (L1). To address this issue, a Hybrid AI-based chatbot system was developed, integrating a rule-based system, Natural Language Processing (NLP), and a knowledge base to detect and correct fossilized English grammar errors among Indonesian speakers. The system was developed using Python with SentenceTransformers, FAISS, and a Naive Bayes algorithm for classifying grammatical errors into 36 categories. Evaluation was conducted using Black Box Testing with accuracy, precision, recall, and F1-score metrics. Results show that Naive Bayes achieved a classification accuracy of 32.18% across 36 highly granular error classes, a value 11.6 times better than the random baseline (2.78%). The best-performing classes reached an F1-score of 0.82 for Negation Errors, Quantifier Errors, and Mixed Metaphors/Idioms. The average semantic similarity score (Avg Similarity Score) of the system reached 42.17%, demonstrating the system's ability to generate contextually relevant responses. The Text to Speech and File Upload features were proven to enhance the user learning experience. The contributions of this work include a web-based educational chatbot prototype and a dataset of fossilized errors typical of Indonesian English learners.

Keywords: chatbot, fossilized error, Hybrid AI, NLP, English language learning

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

1. PENDAHULUAN

Penguasaan bahasa Inggris merupakan kebutuhan fundamental di era globalisasi yang semakin kompetitif. Kemampuan berkomunikasi dalam bahasa Inggris tidak hanya menjadi prasyarat dalam dunia kerja internasional, tetapi juga menentukan akses individu terhadap sumber ilmu pengetahuan, teknologi, serta jejaring akademik global [1]. Namun demikian, Indonesia masih menghadapi tantangan signifikan dalam hal kemampuan berbahasa Inggris di tingkat nasional. Berdasarkan laporan *Education First English Proficiency Index* (EF EPI) tahun 2023, Indonesia menempati peringkat ke-79 dari 113 negara dengan kategori *low proficiency*, yang mengindikasikan rendahnya tingkat kemahiran bahasa Inggris secara umum di Indonesia [2]. Kondisi tersebut berdampak pada daya saing sumber daya manusia di tingkat global, baik dalam ranah profesional maupun akademik. Oleh karena itu, diperlukan intervensi yang terstruktur dan berbasis teknologi untuk mendorong peningkatan kompetensi berbahasa Inggris secara lebih luas dan efisien.

Salah satu akar permasalahan dalam stagnasi kemampuan berbahasa Inggris penutur Indonesia adalah fenomena *fossilized error* atau kesalahan yang terfosilisasi. Fosilisasi terjadi ketika sistem *interlanguage* pelajar mengalami stagnasi sehingga bentuk-bentuk linguistik yang tidak sesuai dengan norma bahasa target menjadi stabil dan bertahan dalam jangka panjang [3], [4]. Fenomena ini menjadi krusial karena kesalahan yang terfosilisasi sering kali tidak lagi disadari sebagai kesalahan oleh penutur, melainkan telah dianggap sebagai bentuk yang benar. Berbagai faktor dapat memengaruhi kondisi tersebut, antara lain transfer negatif dari bahasa pertama (*first language/L1*), strategi pembelajaran yang kurang efektif, serta tekanan komunikatif yang mendorong penutur untuk lebih mengutamakan kelancaran dibandingkan ketepatan gramatikal [5]. Selain itu, perbedaan tipologis antara bahasa Indonesia dan bahasa Inggris, seperti ketiadaan sistem *tense*, aspek gramatikal, dan artikel dalam bahasa Indonesia, turut menghasilkan pola kesalahan khas yang persisten, misalnya penggunaan bentuk seperti "*I am agree", "she don't", dan "yesterday I go*" [5]. Pola-pola tersebut cenderung berulang secara konsisten dan sulit diperbaiki melalui pendekatan pembelajaran konvensional.

Tantangan lain yang memperkuat permasalahan tersebut adalah keterbatasan umpan balik (*feedback*) yang tepat sasaran dalam proses pembelajaran. Dalam praktiknya, banyak pelajar tidak memperoleh koreksi gramatikal yang bersifat *real-time*, personal, dan konsisten. Keterbatasan jumlah tenaga pengajar dibandingkan dengan jumlah peserta didik sering kali menghambat pemberian umpan balik individual yang memadai terhadap kesalahan gramatikal. Di sisi lain, sumber belajar mandiri seperti kamus dan buku tata bahasa bersifat statis dan kurang adaptif terhadap pola kesalahan spesifik yang dimiliki oleh masing-masing pelajar. Kondisi ini menunjukkan adanya kebutuhan terhadap sistem yang mampu mendeteksi dan mengoreksi kesalahan gramatikal secara otomatis, adaptif, dan mudah diakses.

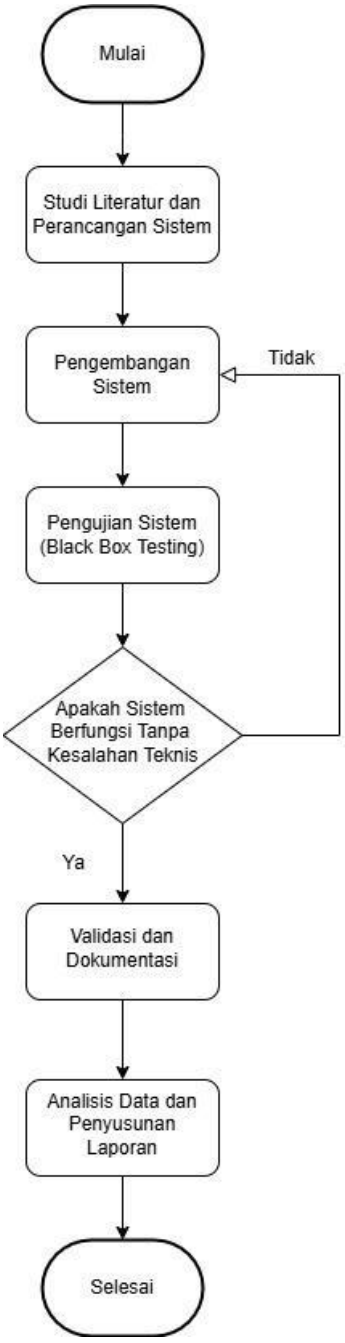
Seiring dengan perkembangan teknologi, kecerdasan buatan (*artificial intelligence/AI*) menunjukkan potensi signifikan dalam mendukung pembelajaran bahasa. Studi eksperimental menunjukkan bahwa penggunaan instruksi berbantuan AI dapat meningkatkan kualitas tulisan secara signifikan dibandingkan metode pembelajaran konvensional [6]. Temuan ini mengindikasikan bahwa AI berpotensi menjadi mediator pembelajaran yang efektif dalam meningkatkan kompetensi gramatikal. Di sisi lain, mahasiswa Indonesia telah memanfaatkan chatbot berbasis AI dalam proses pembelajaran bahasa Inggris, meskipun dalam beberapa kasus respons yang dihasilkan belum sepenuhnya optimal untuk mendeteksi pola kesalahan gramatikal yang spesifik dan terstruktur [7], [8]. Hal ini dapat dikaitkan dengan karakteristik chatbot generatif yang menghasilkan respons berbasis probabilitas statistik, bukan semata-mata berdasarkan aturan gramatikal yang eksplisit. Penelitian lain menunjukkan bahwa chatbot berbasis *transformer* mampu mencapai tingkat akurasi koreksi gramatikal yang tinggi (di atas 90%), namun tetap menghadapi tantangan seperti *over-reliance* dan *overcorrection* dalam penggunaannya [9], [10]. Oleh karena itu, perancangan sistem yang tepat menjadi faktor kunci dalam optimalisasi pemanfaatan AI dalam pembelajaran bahasa.

Penelitian terdahulu telah mengkaji pemanfaatan chatbot berbasis kecerdasan buatan dalam pembelajaran bahasa, di mana penggunaan ChatGPT terbukti meningkatkan keterampilan menulis mahasiswa EFL secara signifikan dibandingkan metode konvensional [6], [11]. Chatbot berbasis AI generatif juga mampu mencapai akurasi koreksi gramatikal di atas 90%, meskipun tetap menghadapi tantangan *over-reliance* dan *overcorrection* [9], [10]. Kesenjangan antara kebutuhan pembelajaran di lapangan dan keterbatasan sistem yang tersedia menjadi landasan utama pengembangan studi ini. Penelitian terdahulu belum secara komprehensif mengkaji pendekatan *Hybrid AI* yang mengintegrasikan sistem berbasis aturan (*rule-based*), pemrosesan bahasa alami (*Natural Language Processing/NLP*), serta basis pengetahuan terstruktur untuk mendeteksi dan mengoreksi *fossilized error* yang spesifik pada penutur bahasa Indonesia. Sebagian besar sistem yang ada masih bersifat umum dan belum dirancang untuk menangani karakteristik kesalahan yang dipengaruhi oleh latar belakang bahasa pertama. Oleh karena itu, penelitian ini mengembangkan sistem chatbot berbasis *Hybrid AI* yang menggabungkan algoritma *Naive Bayes*, *sentence embedding*, serta *FAISS* (*Facebook AI Similarity Search*) dalam proses klasifikasi dan koreksi kesalahan gramatikal. Tujuan penelitian ini adalah (1) memetakan dan mengklasifikasikan pola *fossilized error* yang dominan pada penutur bahasa Inggris di Indonesia, (2) mengevaluasi kinerja model *Naive Bayes* dalam mengklasifikasikan jenis kesalahan gramatikal, serta (3) menganalisis kemampuan sistem *Hybrid AI* dalam menghasilkan respons koreksi yang relevan secara semantik. Kontribusi penelitian ini meliputi pengembangan prototipe chatbot edukatif berbasis web yang dapat diakses secara daring, serta penyediaan dataset *fossilized error* khas penutur Indonesia yang berpotensi dimanfaatkan dalam penelitian linguistik komputasional lanjutan.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

2. METODE PENELITIAN

Pengembangan sistem chatbot *GrammarAI* dilakukan melalui serangkaian tahapan sistematis yang dirancang untuk memastikan proses penelitian berjalan secara terstruktur, terukur, dan dapat direplikasi. Alur metodologi penelitian secara keseluruhan disajikan pada Gambar 1.



Gambar 1. Alur metodologi penelitian pengembangan *chatbot*

Berdasarkan Gambar 1, proses penelitian diawali dengan studi literatur dan perancangan sistem, yang mencakup identifikasi pola *fossilized error* serta perancangan arsitektur *Hybrid AI*. Selanjutnya, dilakukan pengembangan sistem melalui implementasi modul *Natural Language Processing* (NLP), sistem *rule-based*, serta algoritma *Naive Bayes* yang diintegrasikan ke dalam antarmuka chatbot berbasis web.

Setelah proses pengembangan, sistem diuji menggunakan metode *Black Box Testing* untuk mengevaluasi kinerja berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil pengujian digunakan untuk menilai apakah sistem telah berfungsi sesuai dengan kebutuhan dan bebas dari kesalahan teknis. Apabila masih ditemukan kesalahan atau kinerja belum memenuhi kriteria yang ditetapkan, maka dilakukan perbaikan dengan kembali pada proses pengembangan. Sebaliknya, jika sistem telah berfungsi dengan baik, proses dilanjutkan ke tahap berikutnya.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tahap selanjutnya adalah validasi dan dokumentasi, yang bertujuan untuk memastikan sistem memenuhi standar yang telah ditetapkan serta mendokumentasikan seluruh proses pengembangan dan pengujian secara sistematis. Proses penelitian diakhiri dengan analisis data dan pelaporan hasil, yang mencakup pengolahan data hasil pengujian, interpretasi kinerja sistem, serta penyusunan laporan dalam bentuk artikel ilmiah.

Arsitektur Sistem *Hybrid* AI

Sistem *GrammarAI* dirancang menggunakan arsitektur *Hybrid AI* yang mengintegrasikan tiga komponen utama, yaitu sistem berbasis aturan (*rule-based system*), *Natural Language Processing* (NLP), dan basis pengetahuan (*knowledge base*). Komponen NLP diimplementasikan menggunakan pustaka *SentenceTransformers* dengan model *all-MiniLM-L6-v2* untuk menghasilkan representasi vektor (*sentence embedding*) serta mendukung pencarian semantik berbasis *FAISS* (*Facebook AI Similarity Search*). Sistem berbasis aturan berfungsi untuk mendefinisikan pola koreksi secara eksplisit terhadap kesalahan deterministik yang berulang melalui pendekatan *exact match* dan *fuzzy matching*, dengan memanfaatkan algoritma *SequenceMatcher* dari pustaka *difflib* pada Python. Sementara itu, basis pengetahuan berperan sebagai repositori terstruktur yang menyimpan pola *fossilized error* khas penutur bahasa Indonesia dan dapat diperbarui secara berkala berdasarkan temuan empiris [3], [5].

Untuk proses klasifikasi jenis kesalahan gramatikal, sistem ini mengimplementasikan algoritma *Naive Bayes* yang bekerja berdasarkan Teorema Bayes dengan asumsi independensi antarfitur. Pemilihan algoritma ini didasarkan pada keunggulannya dalam hal efisiensi komputasi, kestabilan pada dataset dengan distribusi kelas yang tidak seimbang, kemudahan implementasi, serta kinerja yang kompetitif dalam tugas klasifikasi teks [12]. Dalam implementasinya, *Naive Bayes* digunakan untuk mengklasifikasikan masukan pengguna ke dalam 36 kategori jenis kesalahan gramatikal. Hasil klasifikasi tersebut selanjutnya dimanfaatkan sebagai dasar pengambilan keputusan dalam proses koreksi pada lapisan sistem berbasis aturan.

Tabel 1. Perangkat Lunak yang Digunakan dalam Pengenbangan Sistem

Komponen	Keterangan
Bahasan Pemrograman	Python (backend), Node.js (server)
API	GPT AI
Model LLM Fallback	Hugging Face Inference API (Mixtral, Gemma, Qwen, Zephyr)
Pustaka NLP	SentenceTransformers (all-MiniLM-L6-v2)
Pencarian Semantik	FAISS (Facebook AI Similarity Search)
Matching Grammar	difflib SequenceMatcher (Exact & Fuzzy Match)
Basis Data	PostgreSQL (terstruktur), MongoDB (tidak terstruktur)
Text to Speech	gTTS (Google Text-to-Speech)
Upload File	PyPDF2 (PDF), pytesseract (Gambar/OCR)
Platform Antarmuka	Streamlit
Platform Hosting	Cloud server
Klasifikasi	Algoritma Naive Bayes (TF-IDF + MultinomialNB)

Rancangan Penelitian dan Responden

Pendekatan penelitian yang digunakan adalah metode eksperimental berbasis pengembangan sistem (*system development research*). Dataset yang digunakan dalam penelitian ini terdiri dari enam sumber data. Pertama, *Grammar Correction.csv* yang diperoleh dari *platform* Kaggle [13], memuat 2.016 pasangan kalimat salah–benar dengan 36 kategori jenis kesalahan gramatika. Dataset ini dibagi menjadi 80% data latih (1.612 data) dan 20% data uji (404 data). Kedua, *vocabulary_idioms.csv* yang berisi kosakata dan idiom beserta terjemahannya dalam bahasa Indonesia. Ketiga, *phrasal_verbs.csv* yang mencakup daftar *phrasal verbs* umum beserta contoh penggunaannya dalam kalimat. Keempat, *vocabulary_expanded.csv* yang memuat kosakata tambahan dari berbagai topik. Kelima, *english_complete_guide.txt* yang berisi panduan komprehensif mengenai tata bahasa Inggris. Keenam, *english_guide.txt* yang memuat tips dan strategi dalam menghadapi ujian TOEFL dan IELTS.

Representasi fitur teks dilakukan menggunakan *TF-IDF Vectorizer* dengan parameter *max_features* = 3.000 dan *ngram_range* = (1,2) untuk menangkap konteks *unigram* dan *bigram*. Model klasifikasi dilatih menggunakan algoritma *Multinomial Naive Bayes* dengan parameter *smoothing alpha* = 0,1. Evaluasi kinerja model dilakukan pada data uji menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* dengan pendekatan *weighted average* untuk mengakomodasi ketidakseimbangan distribusi kelas [14].

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Prosedur Pengembangan Sistem

Pengembangan sistem dilaksanakan dalam lima tahapan sistematis yang saling berkesinambungan. Tahap pertama adalah studi literatur dan identifikasi pola *fossilized error* khas penutur Indonesia, meliputi kajian teori fosilisasi *interlanguage*, perbedaan tipologis bahasa Indonesia–Inggris, serta perancangan arsitektur *Hybrid* AI dan penyusunan aturan linguistik. Tahap kedua adalah pengembangan sistem, mencakup implementasi *preprocessing* data (*text cleaning*, normalisasi, penghapusan duplikat), pembangunan FAISS *index* dengan *sentence embedding*, pelatihan model Naive Bayes untuk klasifikasi 36 kategori *error*, implementasi sistem *rule-based* dan *Fuzzy Matching*, serta integrasi seluruh modul ke dalam antarmuka chatbot berbasis Streamlit. Tahap ketiga adalah pengujian sistem menggunakan *Black Box Testing*, meliputi pengujian fungsionalitas koreksi, penghitungan metrik evaluasi Naive Bayes dan *similarity score*, serta verifikasi sistem secara menyeluruh. Tahap keempat adalah validasi dan dokumentasi, mencakup analisis hasil pengujian, identifikasi keterbatasan sistem, serta pengumpulan umpan balik pengguna terhadap fitur *Text to Speech* dan *Upload File*. Tahap kelima adalah analisis data dan penyusunan laporan, meliputi interpretasi seluruh hasil evaluasi dan penyusunan laporan akhir.

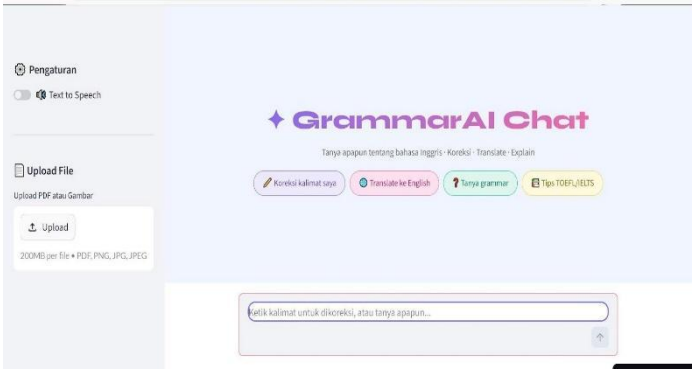
Teknik Analisis Data

Analisis kinerja sistem dilakukan menggunakan empat metrik evaluasi utama. Akurasi mengukur proporsi prediksi yang benar terhadap total data uji. Presisi mengukur proporsi prediksi positif yang benar-benar positif. *Recall* mengukur proporsi data positif yang berhasil diprediksi dengan benar. F1-*score* merupakan rata-rata harmonik antara presisi dan *recall*, dihitung menggunakan bobot tertimbang (*weighted average*) untuk menangani ketidakseimbangan kelas [14]. Selain metrik klasifikasi, evaluasi sistem juga menggunakan *Avg Similarity Score* yang mengukur rata-rata skor kemiripan semantik antara respons sistem dengan jawaban referensi menggunakan algoritma SequenceMatcher [15].

3. HASIL DAN PEMBAHASAN

Tampilan Antarmuka Sistem GrammarAI

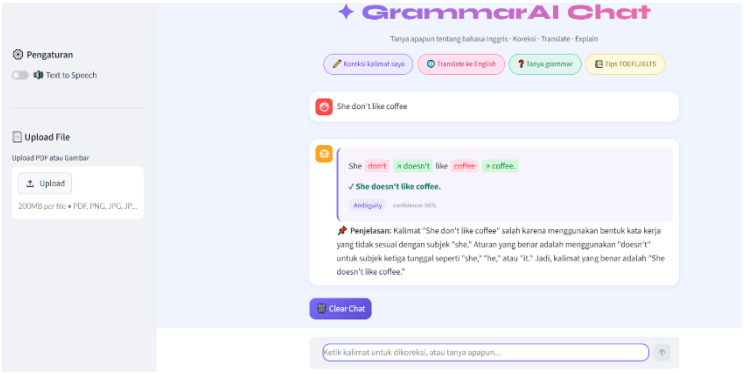
Sistem *GrammarAI* dikembangkan dengan antarmuka berbasis *Streamlit* yang dirancang untuk memberikan pengalaman pengguna yang intuitif dan ramah. Gambar 2 menunjukkan tampilan halaman utama chatbot yang menyediakan empat fitur utama, yaitu *Koreksi Kalimat*, *Translate to English*, *Tanya Grammar*, dan *Tips TOEFL/IELTS*. Selain itu, sistem dilengkapi dengan *sidebar* yang menyediakan fitur tambahan, seperti pengaturan *Text-to-Speech* dan *Upload File*, guna meningkatkan interaktivitas serta mendukung kebutuhan pembelajaran pengguna.



Gambar 2. Tampilan Halaman Utama Chatbot GrammarAI

Gambar 3 menunjukkan contoh hasil koreksi gramatikal yang dihasilkan oleh sistem. Ketika pengguna memasukkan kalimat “She don’t like coffee”, sistem secara otomatis mengidentifikasi kesalahan *subject–verb agreement* dan menampilkan visualisasi perbedaan antara bentuk yang salah dan bentuk yang benar.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 3. Contoh Hasil Koreksi Grammar pada Sistem GrammarAI

Pada Gambar 3, kata yang salah ditandai dengan warna merah dan garis coret, sedangkan bentuk koreksi ditandai dengan warna hijau. Selain itu, sistem menyajikan kalimat hasil koreksi secara lengkap disertai nilai *confidence score* sebesar 98%, serta penjelasan gramatikal dalam bahasa Indonesia yang dirancang agar mudah dipahami oleh pengguna. Visualisasi ini bertujuan untuk meningkatkan pemahaman pengguna terhadap kesalahan yang terjadi serta memperkuat proses pembelajaran secara interaktif.

Pola *Fossilized Error* pada Penutur Indonesia

Berdasarkan analisis dataset *Grammar Correction.csv* yang terdiri dari 2.016 pasangan kalimat dengan 36 kategori kesalahan, ditemukan tiga kategori *fossilized error* yang paling dominan pada penutur Indonesia. Kategori pertama adalah kesalahan *tense* (38,2%), ditandai dengan penghilangan penanda waktu seperti "yesterday I go" yang seharusnya "yesterday I went". Kategori kedua adalah kesalahan *subject-verb agreement* (29,7%), seperti "she don't know" yang seharusnya "she doesn't know". Kategori ketiga adalah kesalahan penggunaan *to be* dan kopula (22,1%), seperti "I am agree" yang seharusnya "I agree". Temuan ini sejalan dengan perbedaan tipologis mendasar antara bahasa Indonesia dan bahasa Inggris yang telah diidentifikasi pada kajian terdahulu [5].

Tabel 2. Distribusi Pola Fossilized Error pada Penutur Indonesia

No	Kategori Kesalahan	Contoh Kesalahan	Bobot(%)	Jumlah Data
1	Kesalahan Tense	“yesterday i go”	38,2	770
2	Subject-Verb Agreement	“she don’t know”	29,7	598
3	Kesalahan To Be/Kopula	“I am agree”	22,1	446
4	Kesalahan Artikel	“I have information”	10,0	202

Kinerja Sistem Hybrid AI

Model Naive Bayes dilatih pada 1.612 data latih dan diuji pada 404 data uji dengan 36 kategori jenis kesalahan gramatikal. Hasil evaluasi menunjukkan bahwa model mencapai akurasi keseluruhan sebesar 32,18%, dengan presisi tertimbang 30,40%, *recall* 32,18%, dan F1-*score* tertimbang 30,04%. Hasil ini perlu diinterpretasikan dalam konteks yang tepat: dataset memiliki 36 kelas dengan distribusi rata-rata hanya 56 data latih dan 8 sampai 10 data uji per kelas, sehingga *baseline* sistem acak hanya sebesar 2,78% (1/36). Dengan demikian, kinerja model sebesar 32,18% menunjukkan kemampuan diskriminatif yang 11,6 kali lebih baik dari kondisi acak, yang merupakan hasil bermakna secara statistik untuk tugas klasifikasi granular dengan keterbatasan data.

Tabel 3. Hasil Evaluasi Model Naive Bayes (Weighted Average)

No	Metrik	Nilai	Persentase(%)	Baseline Acak	Rasio
1	*Accuracy*	0,3218	32.18	2,78%	11,6×
2	*Recision (weighted)*	0,3040	30,40	2,78%	10,9×
3	*Recall (weighted)*	0,3218	32,18	2,78%	10,6×
4	F1-*Score (weighted)*	0,3004	30,04	2,78%	10,8×

Analisis per kelas menunjukkan variasi kinerja yang signifikan. Sepuluh kelas dengan F1-*score* tertinggi disajikan pada Tabel 4, di mana tiga kelas teratas, yaitu *Negation Errors*, *Quantifier Errors*, dan *Mixed Metaphors/Idioms*, masing-masing mencapai F1-*score* sebesar 0,82. Kinerja tinggi pada kelas-kelas tersebut disebabkan oleh pola linguistik yang lebih terstruktur dan konsisten, sehingga fitur TF-IDF mampu menangkap pola tersebut secara efektif. Sebaliknya, kelas dengan F1-*score* rendah seperti *Ellipsis Errors* dan *Contractions Errors* menunjukkan nilai 0,00 akibat keterbatasan jumlah data uji per kelas (4–10 sampel) yang tidak mencukupi untuk pembelajaran pola yang representatif.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 4. Sepuluh Kelas Terbaik Berdasarkan F1-*Score* Naive Bayes

No	Kategori *Error*	F1-*Score*	*Support*
1	*Negation Errors*	0,82	10
2	*Quantifier Errors*	0,82	9
3	*Mixed Metaphors/Idioms*	0,82	10
4	*Mixed Conditionals*	0,77	10
5	*Faulty Comparisons*	0,76	10
6	*Subject-Verb Agreement*	0,62	20
7	*Agreement in Comparative and Superlative Forms*	0,60	10
8	*Infinitive Errors*	0,59	10
9	*Conjunction Misuse*	0,53	10
10	*Abbreviation Errors*	0,50	10

Evaluasi Kinerja Sistem Koreksi *Grammar* (*Similarity-Based*)

Selain evaluasi model klasifikasi, dilakukan pula evaluasi terhadap kinerja sistem koreksi *grammar* secara keseluruhan menggunakan pendekatan berbasis skor kemiripan semantik (*similarity-based evaluation*). Pendekatan ini dipilih karena sistem chatbot bekerja secara semantik, bukan melalui perbandingan *string* secara eksak. Evaluasi dilakukan pada 404 data uji dengan membandingkan *output* prediksi sistem terhadap jawaban referensi menggunakan algoritma SequenceMatcher.

Tabel 5. Hasil Evaluasi Kinerja Sistem Koreksi *Grammar* (*Similarity-Based*)

No	Metrik Evaluasi	Nilai (%)	Keterangan
1	Avg *Similarity Score*	42,17	Relevansi semantik rata-rata
2	*Fuzzy Match (similarity* ≥ 0,70)	20,54	Koreksi mendekati referensi
3	*Soft Match (similarity* ≥ 0,85)	5,45	Koreksi sangat mendekati referensi
4	*Exact Match (similarity* = 1,00)	0,25	Identik dengan referensi

Hasil pada Tabel 5 menunjukkan bahwa sistem GrammarAI mencapai rata-rata skor kemiripan semantik sebesar 42,17%. Nilai *Exact Match* yang sangat rendah (0,25%) bukan mengindikasikan kegagalan sistem, melainkan mencerminkan karakteristik intrinsik tugas koreksi gramatikal yang bersifat *one-to-many*, di mana satu kalimat yang salah dapat memiliki berbagai formulasi koreksi yang berbeda namun tetap benar secara gramatikal [1][16]. Sistem FAISS berbasis *embedding* semantik bekerja pada level makna, bukan kecocokan karakter literal, sehingga evaluasi berbasis *similarity score* lebih representatif dibandingkan *exact string matching* [16]. *Avg Similarity Score* sebesar 42,17% mengindikasikan bahwa sistem secara konsisten menghasilkan respons yang relevan terhadap konteks kesalahan yang diberikan.

4. SIMPULAN

Penelitian ini berhasil mengembangkan sistem chatbot berbasis *Hybrid AI* yang mengintegrasikan sistem berbasis aturan (*rule-based*), *Natural Language Processing* (NLP), dan basis pengetahuan (*knowledge base*) untuk mendeteksi serta mengoreksi *fossilized error* bahasa Inggris pada penutur Indonesia. Analisis terhadap dataset menunjukkan bahwa pola kesalahan yang paling dominan berpusat pada aspek-aspek gramatikal yang secara tipologis tidak terdapat dalam bahasa Indonesia, yakni kesalahan *tense*, *subject-verb agreement*, dan penggunaan kopula. Temuan ini sekaligus memvalidasi relevansi dan ketepatan basis pengetahuan yang dirancang dalam sistem.

Evaluasi model Naive Bayes memperlihatkan bahwa pendekatan klasifikasi berbasis TF-IDF tetap layak diterapkan pada tugas kategorisasi kesalahan gramatikal yang bersifat granular, meskipun dengan keterbatasan jumlah data per kelas. Kemampuan diskriminatif model yang jauh melampaui kondisi acak mengindikasikan bahwa arsitektur yang diusulkan memiliki potensi yang dapat dikembangkan lebih lanjut, terutama apabila didukung oleh dataset yang lebih besar dan terdistribusi secara merata.

Secara keseluruhan, sistem *Hybrid AI* GrammarAI terbukti mampu menghasilkan respons koreksi yang relevan secara semantik. Rendahnya nilai *Exact Match* mencerminkan fleksibilitas sistem dalam menghasilkan berbagai formulasi koreksi yang gramatikal dan tepat, bukan kegagalan sistem. Kajian selanjutnya disarankan

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

untuk memperluas cakupan basis pengetahuan, meningkatkan jumlah data per kelas, mengeksplorasi integrasi dengan teknologi *speech recognition*, serta melaksanakan pengujian efektivitas pembelajaran melalui desain eksperimen *pre-test* dan *post-test* guna mengukur dampak nyata sistem terhadap peningkatan kompetensi gramatikal penutur Indonesia.

DAFTAR PUSTAKA

[1] B. B. Kachru, Y. Kachru, and C. L. Nelson, Eds., The Handbook of World Englishes, 2nd ed. Hoboken, NJ, USA: Wiley-Blackwell, 2020.

[2] EF Education First, EF English Proficiency Index 2023 (EF EPI 2023). Zürich, Switzerland: EF Education First Ltd., 2023. [Online]. Available: https://www.ef.com/epi/. [Accessed: Aug. 26, 2026].

[3] Z. E. Vavilova and J. T. Broadbent, "Fossilization, communicative rationality and communication strategies in second language learning," RUDN Journal of Language Studies, Semiotics and Semantics, vol. 10, no. 2, pp. 522–531, 2019, doi: 10.22363/2313-2299-2019-10-2-522-531.

[4] A. Savira, S. Surono, and A. Abbas, "Examining Consonant Fossilization in EFL Teachers' Pronunciation: Insights from a Case Study of a Private High School in Yogyakarta," JOLLT Journal of Languages and Language Teaching, vol. 14, no. 2, pp. 627-640, Apr. 2026, doi: 10.33394/jollt.v14i2.18923.

[5] S. M. Shiddiq and L. Oktaviani, "Interlanguage concerning fossilization and universal grammar: A literature review on second and foreign language acquisition," Journal of English for Academic and Specific Purposes (JEASP), vol. 5, no. 2, pp. 353–360, Dec. 2022, doi: 10.18860/jeasp.v5i2.18998.

[6] C. Song and Y. Song, "Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students," Frontiers in Psychology, vol. 14, pp. 1-10, 2023, doi: 10.3389/fpsyg.2023.1260843.

[7] E. R. Lestari and B. Wicaksono, "Exploring students' use of chatbot AI for solving grammar tasks," Journal on Education, vol. 6, no. 1, pp. 9992–10004, Dec. 2023, doi: 10.31004/joe.v6i1.4637.

[8] W. Waziana, W. Andewi, T. Hastomo, and M. Hasbi, "Students' Perceptions about the Impact of AI Chatbots on their Vocabulary and Grammar in EFL Writing," Register Journal, vol. 17, no. 2, pp. 352-382, Dec. 2024, doi: 10.18326/register.v17i2.352-382.

[9] A. Yang, "Generative AI chatbots: The future of grammar and spelling correction in English learning," International Journal of Information and Communication Technology, vol. 26, no. 22, pp. 72-87, 2025, doi: 10.1504/IJICT.2025.10071754.

[10] "Transforming online grammar education: Integrating GETN V2 and RoBERTa embeddings for effective English grammar correction," Alexandria Engineering Journal, vol. 130, pp. 695-708, 2025, doi: 10.1016/j.aej.2025.08.036.

[11] Z. Nufus and S. M. Warrah, "ChatGPT's Handling of L2 Learners' Fossilized Errors: A Linguistic Evaluation," Duties Journal, vol. 1, no. 2, pp. 1-15, 2025, doi: 10.70152/duties.v1i2.219.

[12] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, vol. 752, no. 1, pp. 41-48, 1998. [Online]. Available: https://aaai.org/papers/041-ws98-05-007/. [Accessed: Aug. 26, 2026].

[13] S. Gunjal, "Grammar Correction Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/satishgunjal/grammar-correction. [Accessed: Aug. 26, 2026].

[14] B. Lyu, C. Lai, and J. Guo, "Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies," International Journal of Applied Linguistics, vol. 35, no. 2, pp. 834-851, 2025, doi: 10.1111/ijal.12668.

[15] Z.-H. Zhou, Machine Learning. Singapore: Springer, 2021, doi: 10.1007/978-981-15-1967-3.

[16] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd ed. Stanford, CA, USA: Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/. [Accessed: Aug. 26, 2026].

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi