E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026

Perbandingan Efektivitas *Back Translation* dan *Easy Data Augmentation* pada *Automatic Short Answer Scoring* Bahasa Indonesia

[1*]Nur Fadilah, [2]Khawaritzmi Abdallah Ahmad, [3]Muh. Isbar Pratama

[1,2,3]Universitas Negeri Makassar

Email: nurfadilah@unm.ac.id[1*], khawaritzmi.abdallah@unm.ac.id[2], isbarpratama@unm.ac.id[3]

Received : 04 Januari 2026
Accepted : 19 Februari 2026
Published : 30 Maret 2026

ABSTRAK

Automatic Short Answer Scoring (AES) merupakan salah satu penerapan *Natural Language Processing* (NLP) yang digunakan untuk melakukan penilaian otomatis terhadap jawaban uraian singkat. Salah satu tantangan utama dalam pengembangan sistem AES adalah keterbatasan jumlah dan keragaman dataset yang dapat memengaruhi kemampuan generalisasi model. Penelitian sebelumnya menunjukkan bahwa metode *Easy Data Augmentation* (EDA) berbasis Sinonim IndoBERT mampu meningkatkan performa model pada dataset UKARA, namun masih memiliki keterbatasan karena augmentasi dilakukan pada tingkat kata. Penelitian ini bertujuan membandingkan efektivitas metode *Back Translation* dan EDA Sinonim IndoBERT pada sistem AES Bahasa Indonesia menggunakan dataset UKARA. Penelitian dilakukan dengan mempertahankan dataset, tahapan *preprocessing*, representasi teks menggunakan FastText, arsitektur BiLSTM, dan metode evaluasi yang sama, sehingga perbedaan performa yang diperoleh hanya dipengaruhi oleh metode augmentasi. Pengujian dilakukan menggunakan skenario *Non K-Fold Evaluation* dan *3-Fold Cross Validation*. Hasil penelitian menunjukkan bahwa *Back Translation* menghasilkan performa yang lebih baik pada sebagian besar skenario pengujian, dengan akurasi tertinggi sebesar 89,00% pada Dataset A. Hasil penelitian juga menunjukkan bahwa kualitas variasi data yang dihasilkan memiliki pengaruh yang lebih besar terhadap performa model dibandingkan sekadar peningkatan jumlah data. Dengan demikian, *Back Translation* dapat menjadi alternatif yang efektif untuk meningkatkan kualitas dataset dan performa sistem AES Bahasa Indonesia.

Kata kunci: *Automatic Short Answer Scoring*, *Back Translation*, *Easy Data Augmentation*, IndoBERT, BiLSTM, FastText.

ABSTRACT

Automatic Short Answer Scoring (AES) is a Natural Language Processing (NLP) application designed to automatically assess short-answer responses. One of the primary challenges in developing AES systems is the limited size and diversity of available datasets, which can adversely affect a model's generalization capability. Previous studies have demonstrated that Easy Data Augmentation (EDA) based on IndoBERT-generated synonyms can improve model performance on the UKARA dataset; however, this approach remains limited because the augmentation process is performed at the word level. This study aims to compare the effectiveness of Back Translation and IndoBERT-based Synonym EDA for Indonesian AES systems using the UKARA dataset. To ensure a fair comparison, the dataset, preprocessing procedures, FastText-based text representation, BiLSTM architecture, and evaluation methods were kept consistent across experiments, allowing performance differences to be attributed solely to the augmentation techniques. The experiments were conducted using both Non-K-Fold Evaluation and 3-Fold Cross-Validation scenarios. The results indicate that Back Translation outperformed IndoBERT-based Synonym EDA in most experimental settings, achieving the highest accuracy of 89.00% on Dataset A. Furthermore, the findings suggest that the quality and semantic diversity of the generated data have a greater impact on model performance than merely increasing the amount of training data. Therefore, Back Translation can serve as an effective alternative for enhancing dataset quality and improving the performance of Indonesian AES systems.

Keywords: *Automatic Short Answer Scoring, Back Translation, Easy Data Augmentation, IndoBERT, BiLSTM, FastText.*

1. PENDAHULUAN

Tes uraian singkat merupakan salah satu bentuk evaluasi yang banyak digunakan dalam proses pembelajaran untuk mengukur tingkat pemahaman peserta didik terhadap suatu materi secara lebih mendalam. Berbeda dengan soal objektif, tes uraian memungkinkan peserta didik menyampaikan jawaban menggunakan susunan kalimat dan pilihan kata yang beragam. Keberagaman tersebut memberikan peluang bagi pendidik untuk menilai kemampuan analisis, pemahaman konsep, dan kemampuan berpikir kritis peserta didik secara lebih komprehensif. Namun demikian, proses penilaian tes uraian memerlukan waktu yang relatif lama dan rentan terhadap subjektivitas penilai, terutama ketika jumlah peserta yang dinilai semakin besar. Dalam hal ini, pemanfaatan teknologi penilaian otomatis (*Automatic Short Answer Scoring*/AES) berbasis kecerdasan buatan dapat membantu menyelaraskan objektivitas penilaian dan mempercepat proses evaluasi secara signifikan dibandingkan penilaian manual oleh manusia.

Perkembangan *Artificial Intelligence* (AI) dan *Natural Language Processing* (NLP) mendorong lahirnya berbagai sistem penilaian otomatis yang mampu membantu proses evaluasi jawaban peserta didik secara lebih cepat dan konsisten. Salah satu implementasi NLP dalam bidang pendidikan adalah *Automatic Short Answer Scoring* (AES), yaitu sistem yang dirancang untuk melakukan penilaian otomatis terhadap jawaban uraian singkat berdasarkan tingkat kesesuaian jawaban dengan kriteria yang telah ditentukan. Dalam konteks Bahasa Indonesia, Herwanto dkk. (2018) mengembangkan sistem UKARA sebagai salah satu sistem AES yang memanfaatkan teknik NLP dan *machine learning* untuk mengklasifikasikan jawaban peserta berdasarkan kesesuaian makna dengan jawaban acuan. Sistem tersebut menunjukkan bahwa pendekatan berbasis NLP memiliki potensi yang besar dalam mendukung proses penilaian otomatis pada tes uraian singkat.

Meskipun demikian, pengembangan model AES masih menghadapi berbagai tantangan, salah satunya adalah keterbatasan jumlah dan keragaman dataset yang digunakan dalam proses pelatihan model. Pada tes uraian singkat, peserta didik dapat menyampaikan jawaban yang benar menggunakan berbagai variasi kata, frasa, maupun struktur kalimat yang berbeda. Akibatnya, model memerlukan data pelatihan yang cukup beragam agar mampu mempelajari berbagai pola bahasa yang muncul dalam jawaban peserta. Zhang et al. (2021) menjelaskan bahwa jumlah dan keragaman data merupakan faktor penting yang memengaruhi kemampuan generalisasi model NLP. Dataset yang terbatas berpotensi menyebabkan model mengalami kesulitan dalam mengenali variasi jawaban baru sehingga performa prediksi menjadi kurang optimal.

Salah satu pendekatan yang banyak digunakan untuk mengatasi keterbatasan dataset adalah augmentasi data. Teknik ini bertujuan menghasilkan data baru secara otomatis dengan tetap mempertahankan informasi penting yang terdapat pada data asli. Dalam penelitian *Automatic Essay Scoring* menggunakan dataset UKARA, Fadilah dan Priyanta (2022) menerapkan metode *Easy Data Augmentation* (EDA) yang terdiri atas *Synonym Replacement (SR), Random Insertion (RI), Random Deletion (RD), dan Random Swap (RS).* Hasil penelitian menunjukkan bahwa augmentasi data mampu menghasilkan variasi data baru dan meningkatkan performa model pada beberapa skenario pengujian. Namun demikian, hasil yang diperoleh tidak menunjukkan pola yang konsisten pada seluruh dataset. Pada Dataset B, metode EDA sinonim IndoBERT mampu meningkatkan performa model hingga mencapai akurasi terbaik sebesar 70,16%, begitu juga pada Dataset A performa terbaik terjadi pada metode EDA Sinonim IndoBERT dengan akurasi sebesar 82,83%. Berdasarkan hasil tersebut, EDA Sinonim IndoBERT dipilih sebagai metode pembanding dalam penelitian ini karena merupakan teknik augmentasi yang menghasilkan performa terbaik pada penelitian sebelumnya menggunakan dataset UKARA. Temuan tersebut menunjukkan bahwa augmentasi data memiliki kontribusi positif dalam meningkatkan performa model melalui penambahan variasi data pelatihan.

Meskipun demikian, metode EDA masih memiliki beberapa keterbatasan karena proses augmentasi dilakukan pada tingkat kata. Wei dan Zou (2019) menjelaskan bahwa EDA bekerja melalui penggantian, penyisipan, penghapusan, dan pertukaran kata secara acak untuk menghasilkan data baru. Meskipun mampu meningkatkan jumlah data pelatihan dengan cepat, pendekatan tersebut berpotensi

menghasilkan kalimat yang kurang alami atau bahkan mengalami pergeseran makna akibat manipulasi kata secara langsung. Kondisi ini menunjukkan bahwa kualitas data hasil augmentasi menjadi faktor penting yang perlu diperhatikan selain jumlah data yang dihasilkan. Salah satu metode augmentasi yang banyak digunakan dalam penelitian NLP untuk menghasilkan variasi kalimat yang lebih alami adalah *Back Translation*. Metode ini menghasilkan data baru melalui proses penerjemahan teks ke bahasa perantara dan kemudian diterjemahkan kembali ke bahasa asal. Berbeda dengan EDA yang bekerja pada tingkat kata, *Back Translation* menghasilkan variasi kalimat melalui perubahan struktur sintaksis dengan tetap mempertahankan makna utama dari teks sumber. Sennrich et al. (2016) menunjukkan bahwa *Back Translation* mampu meningkatkan kualitas data pelatihan dan memberikan dampak positif terhadap performa model pada berbagai tugas NLP, terutama pada kondisi keterbatasan data.

Meskipun *Back Translation* telah dieksplorasi pada dataset UKARA menggunakan representasi tingkat lanjut berbasis *Sentence-BERT* (SBERT) oleh Tanaka dkk. (2024), efektivitas relatifnya jika dibandingkan secara langsung dengan metode *Easy Data Augmentation* (EDA) di bawah kondisi eksperimen yang identik masih memerlukan kajian mendalam. Khususnya, belum ada penelitian yang mengisolasikan variabel pengklasifikasi dengan membandingkan efektivitas *Back Translation* dan EDA berbasis Sinonim IndoBERT secara langsung menggunakan arsitektur baseline klasik yang sama, yaitu representasi FastText dan model sekuensial klasik *Bidirectional Long Short-Term Memory* (BiLSTM). Perbandingan terkontrol ini sangat penting untuk memahami sejauh mana manipulasi sintaksis (melalui penerjemahan balik) dan manipulasi leksikal (melalui substitusi sinonim kontekstual) memengaruhi performa model sekuensial klasik yang secara komputasi lebih ringan namun sangat sensitif terhadap variabilitas struktural kalimat.

Penelitian ini merupakan pengembangan dari penelitian sebelumnya yang menggunakan metode *Easy Data Augmentation* pada dataset UKARA. Seluruh tahapan penelitian, mulai dari dataset yang digunakan, proses prapemrosesan, representasi teks menggunakan FastText, arsitektur BiLSTM, parameter model, hingga metode evaluasi dipertahankan sama dengan penelitian sebelumnya. Perbedaan utama penelitian ini hanya terletak pada metode augmentasi yang digunakan, yaitu *Back Translation*. Dengan demikian, perbedaan performa yang dihasilkan dapat dianalisis secara lebih objektif sebagai dampak dari penggunaan metode augmentasi yang berbeda.

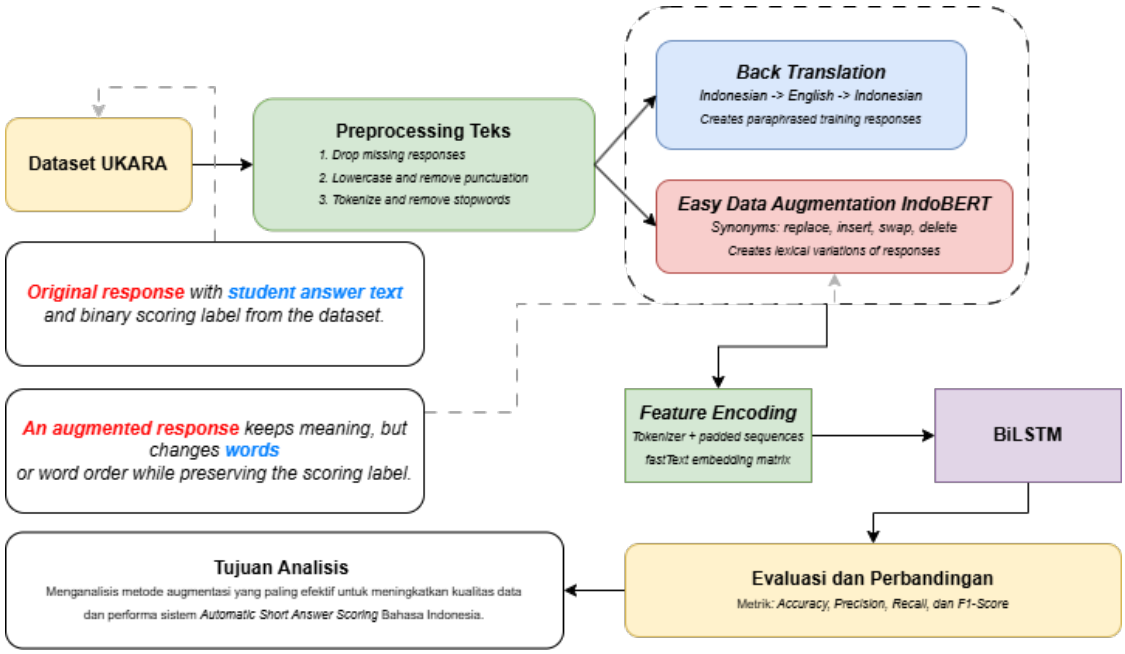
Berdasarkan uraian tersebut, penelitian ini bertujuan untuk membandingkan efektivitas metode *Back Translation* dan *Easy Data Augmentation* pada *Automatic Short Answer Scoring* Bahasa Indonesia menggunakan dataset UKARA. Analisis dilakukan terhadap kualitas data hasil augmentasi serta dampaknya terhadap performa model pada Dataset A and Dataset B. Hasil penelitian ini diharapkan dapat memberikan rekomendasi mengenai metode augmentasi yang lebih efektif dalam meningkatkan kualitas dataset dan performa sistem penilaian otomatis tes uraian singkat Bahasa Indonesia.

2. METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen komparatif untuk membandingkan efektivitas metode *Easy Data Augmentation* (EDA) dan *Back Translation* pada sistem *Automatic Short Answer Scoring* (AES) Bahasa Indonesia. Seluruh tahapan penelitian, mulai dari dataset, proses prapemrosesan, representasi teks menggunakan FastText, arsitektur model BiLSTM, parameter pelatihan, hingga metode evaluasi dipertahankan sama dengan penelitian sebelumnya. Perbedaan utama penelitian ini terletak pada metode augmentasi yang digunakan, yaitu *Easy Data Augmentation* (EDA) dan *Back Translation*. Melalui isolasi variabel kontrol yang ketat ini, perbedaan performa yang diperoleh dapat dianalisis secara lebih objektif sebagai dampak murni dari karakteristik morfologis dan sintaksis yang dihasilkan oleh masing-masing metode augmentasi. Penggunaan metode komparatif yang terisolasi ini sejalan dengan kerangka pengujian performa klasifikasi teks bersumber daya rendah yang direkomendasikan oleh Siahaan dkk. (2023). Alur penelitian yang digunakan pada penelitian ini ditunjukkan pada Gambar 1.

E-ISSN: 2986-7258
P-ISSN : 2986-724X

PISCES Volume 04 Nomor 01 Maret 2026



Gambar 1. Alur Penelitian Perbandingan Metode *Easy Data Augmentation* dan *Back Translation* pada Dataset UKARA

Berdasarkan Gambar 1, penelitian diawali dengan penggunaan dataset UKARA yang terdiri atas Dataset A dan Dataset B sebagai sumber data penelitian. Dataset tersebut kemudian melalui tahap *preprocessing* untuk menghasilkan data yang lebih bersih dan konsisten sebelum dilakukan proses augmentasi. Tahap *preprocessing* meliputi *filtering, case folding*, tokenisasi, dan *padding*. Reduksi derau (*noise*) melalui prapemrosesan teks terstruktur terbukti krusial dalam menstabilkan ekstraksi fitur pada model berbasis sekuensial (Nurhaliza & Al Fatta, 2024).

Setelah tahap *preprocessing*, data diproses menggunakan dua metode augmentasi yang berbeda, yaitu *Easy Data Augmentation* (EDA) Sinonim IndoBERT dan *Back Translation*. Pada metode EDA, proses augmentasi dilakukan menggunakan pendekatan Sinonim IndoBERT untuk menghasilkan variasi kata yang memiliki kesamaan makna dengan teks asli. Sebagai rujukan utama penelitian ini, Fadilah dan Priyanta (2022) membuktikan bahwa penyesuaian leksikal melalui sinonim IndoBERT mampu menyajikan alternatif fraseologis yang fungsional tanpa mengubah orientasi makna asli esai peserta didik. Sementara itu, pada metode *Back Translation*, teks bahasa Indonesia diterjemahkan ke bahasa Inggris sebagai bahasa perantara dan kemudian diterjemahkan kembali ke bahasa Indonesia sehingga menghasilkan variasi kalimat baru dengan struktur sintaksis yang berbeda namun tetap mempertahankan makna utama dari teks sumber. Pemanfaatan *Back Translation* terbukti mampu memperkaya diversifikasi semantik teks sekaligus meningkatkan ketahanan (*robustness*) model secara signifikan terhadap variasi ekspresi baru pada klasifikasi teks Bahasa Indonesia (Rahma & Suadaa, 2023)

Dataset hasil augmentasi kemudian direpresentasikan ke dalam bentuk vektor menggunakan FastText Word Embedding. Representasi vektor tersebut digunakan sebagai masukan bagi model *Bidirectional Long Short-Term Memory* (BiLSTM) untuk melakukan proses klasifikasi jawaban. Sani dan Sarwani (2022) menerapkan representasi kata FastText yang dikombinasikan dengan algoritma pembelajaran saraf untuk mengoreksi jawaban esai berdasarkan persamaan makna semantik, membuktikan efektivitas metode ini dalam menangkap informasi kontekstual yang padat. Karakteristik BiLSTM yang memproses sekuensial dari dua arah dinilai sangat ideal karena mampu menangkap konteks maju dan mundur secara simultan, yang terbukti unggul untuk pemodelan teks pendek berkarakteristik padat (Purwoko dkk., 2023). Penggunaan FastText dan BiLSTM dipertahankan sesuai

penelitian sebelumnya untuk menjaga konsistensi lingkungan eksperimen sehingga pengaruh metode augmentasi dapat dianalisis secara lebih objektif.

Tahap selanjutnya adalah evaluasi model yang dilakukan menggunakan dua skenario pengujian, yaitu *Non K-Fold Evaluation* dan 3-Fold *Cross Validation*. Pada skenario *Non K-Fold*, model dievaluasi menggunakan pembagian data yang telah ditentukan pada dataset UKARA. Sementara itu, pada skenario *3-Fold Cross Validation*, data dibagi menjadi tiga subset yang digunakan secara bergantian sebagai data pelatihan dan data pengujian untuk memperoleh hasil evaluasi yang lebih stabil. Kinerja model diukur menggunakan metrik *Accuracy, Precision, Recall,* dan *F1-Score* yang diperoleh dari *confusion matrix.* Hasil evaluasi dari kedua metode augmentasi kemudian dibandingkan untuk menganalisis efektivitas masing-masing metode dalam meningkatkan kualitas data hasil augmentasi and performa sistem *Automatic Short Answer Scoring* Bahasa Indonesia. Analisis tersebut menjadi dasar dalam menentukan metode augmentasi yang paling efektif untuk digunakan pada dataset UKARA.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset UKARA yang dikembangkan oleh Herwanto dkk. (2018) dan digunakan kembali oleh Fadilah dan Priyanta (2022) untuk penelitian *Automatic Short Answer Scoring* (AES) Bahasa Indonesia. Dataset UKARA terdiri atas jawaban uraian singkat berbahasa Indonesia yang telah diberi label benar dan salah berdasarkan kesesuaiannya dengan jawaban acuan. Pada penelitian ini digunakan dua kelompok dataset, yaitu Dataset A dan Dataset B, yang masing-masing dibagi menjadi data pelatihan (*training*), validasi (*validation*), dan pengujian (*testing*). Distribusi dataset ditunjukkan pada Tabel 1.

Tabel 1. Sebaran Data UKARA

Data	Jumlah Data	Label Benar	Label Salah
Training_A	854	609	245
Validation_A	215	153	62
Testing_A	268	191	77
Training_B	974	531	437
Validation_B	244	135	109
Testing_B	305	168	137

Berdasarkan Tabel 1, Dataset A terdiri atas 1.337 data dengan 953 label benar dan 384 label salah, sedangkan Dataset B terdiri atas 1.523 data dengan 834 label benar dan 683 label salah. Dataset A memiliki distribusi kelas yang lebih tidak seimbang dibandingkan Dataset B yang relatif lebih seimbang. Perbedaan karakteristik tersebut memungkinkan analisis pengaruh metode augmentasi pada kondisi dataset yang berbeda. Untuk menjaga konsistensi lingkungan eksperimen dengan penelitian sebelumnya, data pelatihan dan validasi digunakan dalam proses augmentasi dan pelatihan model, sedangkan data pengujian digunakan sebagai data evaluasi tanpa melalui proses augmentasi.

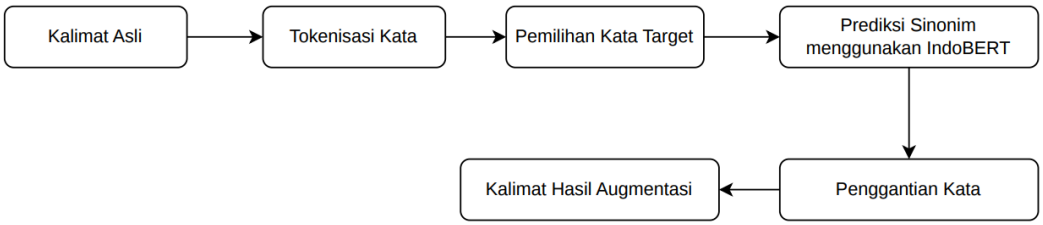
2.2 Prepocessing

Tahap *preprocessing* dilakukan untuk membersihkan dan menyeragamkan data sebelum proses augmentasi dan pelatihan model. Proses ini bertujuan untuk mengurangi noise pada data serta menghasilkan representasi teks yang lebih konsisten. Tahapan *preprocessing* yang digunakan mengacu pada penelitian Fadilah dan Priyanta (2022), yang meliputi *filtering*, *case folding*, *tokenization*, dan *padding*. Filtering digunakan untuk menghapus karakter yang tidak diperlukan, *case folding* dilakukan untuk mengubah seluruh huruf menjadi huruf kecil, *tokenization* digunakan untuk memecah kalimat menjadi token-token kata, sedangkan padding dilakukan untuk menyamakan panjang urutan token agar dapat diproses oleh model BiLSTM. Seluruh tahapan *preprocessing* diterapkan pada Dataset A dan Dataset B sebelum dilakukan proses augmentasi menggunakan EDA Sinonim IndoBERT maupun *Back Translation*.

2.3 *Easy Data Augmentation* (EDA) dengan Sinonim Indobert

Metode *Easy Data Augmentation* (EDA) yang digunakan dalam penelitian ini mengacu pada penelitian Fadilah dan Priyanta (2022). Berdasarkan hasil penelitian tersebut, EDA berbasis Sinonim IndoBERT menghasilkan performa terbaik dibandingkan variasi augmentasi lainnya pada dataset UKARA, dengan akurasi sebesar 82,83% pada Dataset A dan 70,16% pada Dataset B. Oleh karena itu, metode ini dipilih sebagai metode pembanding dalam penelitian ini.

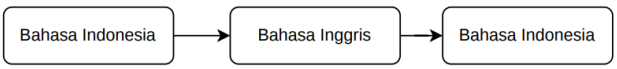
EDA Sinonim IndoBERT bekerja dengan menghasilkan variasi data baru melalui proses penggantian kata menggunakan sinonim yang diperoleh dari model IndoBERT. Berbeda dengan pendekatan berbasis kamus sinonim, IndoBERT memanfaatkan representasi kontekstual sehingga kata pengganti yang dihasilkan lebih sesuai dengan konteks kalimat. Pendekatan ini memungkinkan proses augmentasi menghasilkan variasi data yang lebih relevan secara semantik tanpa mengubah makna utama dari teks sumber. Proses augmentasi dilakukan dengan memilih sejumlah kata dalam kalimat dan menggantinya dengan sinonim yang memiliki tingkat kesamaan makna tertinggi berdasarkan model IndoBERT. Hasil augmentasi kemudian digabungkan dengan data asli untuk membentuk dataset pelatihan yang lebih beragam.



Gambar 2. Proses EDA Sinonim IndoBERT

2.4 Back Translation

Pada penelitian ini, proses *Back Translation* dilakukan dengan menerjemahkan teks berbahasa Indonesia ke dalam Bahasa Inggris sebagai bahasa perantara, kemudian hasil terjemahan tersebut diterjemahkan kembali ke Bahasa Indonesia. Proses ini memungkinkan terbentuknya variasi kalimat baru melalui perubahan struktur sintaksis, pemilihan kosakata, maupun susunan frasa tanpa mengubah informasi utama yang terkandung dalam kalimat. Secara umum, proses *Back Translation* dapat dinyatakan sebagai berikut:



Gambar 3. Proses *Back Translation*

Hasil terjemahan yang diperoleh selanjutnya digunakan sebagai data augmentasi dan digabungkan dengan data asli untuk membentuk dataset pelatihan yang lebih beragam. Dengan pendekatan ini, model diharapkan mampu mempelajari variasi ekspresi bahasa yang lebih luas sehingga meningkatkan kemampuan generalisasi dalam melakukan klasifikasi jawaban uraian singkat.

2.5 Representasi Teks Menggunakan FastText

Representasi teks pada penelitian ini menggunakan model FastText bahasa Indonesia pra-latih (*cc.id.300.bin*) dengan dimensi vektor 300. FastText dipilih karena memanfaatkan representasi *subword* yang memungkinkan pembentukan vektor untuk kata yang tidak terdapat dalam kosakata pelatihan (*Out-of-Vocabulary/OOV*). Kemampuan ini sesuai dengan karakteristik bahasa Indonesia yang memiliki keragaman morfologi dan penggunaan afiks yang tinggi. Selain itu, FastText tetap mampu menghasilkan representasi yang baik untuk kata baru maupun kata yang mengalami kesalahan

penulisan. Ariyus dkk. (2024) menunjukkan bahwa pendekatan ini efektif dalam mempertahankan performa klasifikasi pada teks bahasa Indonesia yang memiliki variasi linguistik yang beragam.

2.6 Arsitektur BiLSTM

Model klasifikasi yang digunakan dalam penelitian ini adalah *Bidirectional Long Short-Term Memory* (BiLSTM). BiLSTM dipilih karena mampu memproses informasi dari dua arah sekaligus, yaitu urutan kata dari awal ke akhir (*forward*) dan dari akhir ke awal (*backward*), sehingga dapat menangkap konteks kalimat secara lebih komprehensif. Pada penelitian ini, model menerima masukan berupa representasi FastText berdimensi 300 yang diperoleh dari proses vektorisasi teks. *Layer embedding* menggunakan bobot FastText pra-latih dan tidak diperbarui selama proses pelatihan (*non-trainable*). Selanjutnya, representasi kata diproses menggunakan *layer* BiLSTM dengan 150 unit untuk mempelajari hubungan kontekstual antar kata. Keluaran dari layer BiLSTM diteruskan ke layer *Dense* berukuran 32 neuron dengan fungsi aktivasi ReLU, kemudian dilakukan regularisasi menggunakan *Dropout* sebesar 0,4 untuk mengurangi risiko *overfitting*. Pada lapisan keluaran digunakan satu *neuron* dengan fungsi aktivasi *sigmoid* untuk menghasilkan prediksi klasifikasi biner berupa label benar atau salah.

2.7 Evaluasi Model

Evaluasi model dilakukan menggunakan dua skenario pengujian, yaitu *Non K-Fold Evaluation* dan *3-Fold Cross Validation.* Pada skenario *Non K-Fold*, model dievaluasi menggunakan pembagian data pelatihan, validasi, dan pengujian yang telah ditetapkan pada dataset UKARA. Sementara itu, pada skenario *3-Fold Cross Validation,* data dibagi menjadi tiga subset yang digunakan secara bergantian sebagai data pelatihan dan data pengujian untuk memperoleh hasil evaluasi yang lebih stabil. Kinerja model diukur menggunakan metrik *Accuracy, Precision, Recall,* dan *F1-Score* yang diperoleh dari *confusion matrix*. Hasil evaluasi dari metode EDA Sinonim IndoBERT dan *Back Translation* kemudian dibandingkan untuk menganalisis efektivitas masing-masing metode augmentasi dalam meningkatkan performa sistem *Automatic Short Answer Scoring* Bahasa Indonesia.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Augmentasi Dataset Pada Skenario *Non K-Fold*

Proses augmentasi dilakukan menggunakan dua metode, yaitu EDA Sinonim IndoBERT dan *Back Translation*. Pada skenario *Non K-Fold Evaluation*, augmentasi diterapkan pada data pelatihan (*training set*), sedangkan pada skenario 3-*Fold Cross Validation* augmentasi diterapkan pada gabungan data pelatihan (*training set*) dan data validasi (*validation set*). Tujuan augmentasi adalah meningkatkan jumlah data sekaligus menyeimbangkan distribusi kelas. Berikut adalah jumlah data hasil augmentasi

Tabel 2. Jumlah Data Hasil Augmentasi pada Skenario *Non K-Fold*

Dataset	Metode	Label Benar	Label Salah	Total
A	EDA Sinonim IndoBERT	609	609	1218
A	*Back Translation*	610	610	1220
B	EDA Sinonim IndoBERT	537	537	1075
B	*Back Translation*	537	537	1074

Berdasarkan Tabel 2, kedua metode augmentasi berhasil menghasilkan distribusi kelas yang seimbang pada data pelatihan. Pada Dataset A, EDA Sinonim IndoBERT menghasilkan 1.218 data, sedangkan *Back Translation* menghasilkan 1.220 data. Sementara itu, pada Dataset B jumlah data hasil augmentasi yang dihasilkan oleh kedua metode relatif sama, yaitu 1.075 data untuk EDA Sinonim IndoBERT dan 1.074 data untuk *Back Translation.*

Tabel 3. Jumlah Data Hasil Augmentasi pada Skenario *3-Fold Cross Validation*

Dataset	Metode	Label Benar	Label Salah	Total
A	EDA Sinonim IndoBERT	762	762	**1524**
A	*Back Translation*	763	763	**1526**
B	EDA Sinonim IndoBERT	672	672	1344
B	*Back Translation*	672	672	1344

Berdasarkan Tabel 3, augmentasi pada skenario *3-Fold Cross Validation* menghasilkan jumlah data yang lebih besar dibandingkan skenario Non K-Fold karena dilakukan pada gabungan data pelatihan dan validasi. Seperti pada skenario sebelumnya, kedua metode berhasil menghasilkan distribusi kelas yang seimbang. Pada Dataset A, jumlah data hasil augmentasi mencapai 1.524 data untuk EDA Sinonim IndoBERT dan 1.526 data untuk *Back Translation*. Sementara itu, pada Dataset B kedua metode menghasilkan jumlah data yang sama, yaitu 1.344 data. Hasil pada Tabel 2 dan Tabel 3 menunjukkan bahwa kedua metode augmentasi mampu mengatasi ketidakseimbangan kelas dan menghasilkan jumlah data yang relatif serupa. Dengan kondisi distribusi kelas yang seimbang, perbedaan performa yang diperoleh pada tahap evaluasi dapat lebih merepresentasikan pengaruh kualitas variasi data yang dihasilkan oleh masing-masing metode augmentasi dibandingkan pengaruh jumlah data atau ketidakseimbangan kelas.

3.2 Analisis Kualitatif Hasil Augmentasi

Selain meningkatkan jumlah data pelatihan, kualitas variasi kalimat yang dihasilkan oleh metode augmentasi juga berperan penting dalam memengaruhi performa model. Oleh karena itu, dilakukan analisis kualitatif terhadap beberapa contoh data hasil augmentasi yang dihasilkan oleh EDA Sinonim IndoBERT dan *Back Translation*. Contoh hasil augmentasi ditunjukkan pada Tabel 4.

Tabel 4. Contoh Hasil Augmentasi Menggunakan EDA Sinonim IndoBERT dan *Back Translation*

Metode	Kalimat Asli	Hasil Augmentasi
EDA Sinonim IndoBERT	mereka perlu menyesuaikan diri dan beradaptasi dengan lingkungan yang baru	mereka perlu **bela diri** dan beradaptasi dengan lingkungan yang baru
EDA Sinonim IndoBERT	beradaptasi dengan lingkungan baru lahan dan harta mereka yang dulu akan hilang tertinggalnya teknologi karena bencana lingkungan tersebut	beradaptasi dengan lingkungan baru lahan dan harta mereka yang dulu akan hilang tertinggalnya **diri** karena bencana lingkungan tersebut
EDA Sinonim IndoBERT	pengungsi akan mencari mata pencaharian yang baru karena mereka kehilangan lahan pertanian mata pencaharian	pengungsi akan mencari mata pencaharian yang baru karena mereka kehilangan lahan **untuk** mata pencaharian
Back Translation	menyumbangkan pakaian	menyumbangkan pakaian
Back Translation	karena sebagai upaya untuk membuat produksi pakaian lebih etis	karena sebagai upaya untuk membuat produksi pakaian lebih etis
Back Translation	sebab pakaian yang mereka pakai tidak cukup; sayang sekali kalau pakaian itu dibuang, lebih baik untuk diberikan	sayang sekali kalau pakaian itu dibuang, lebih baik untuk diberikan

Berdasarkan Tabel 4, metode EDA Sinonim IndoBERT menghasilkan variasi data melalui penggantian kata menggunakan prediksi token dari model bahasa. Meskipun pendekatan ini mampu menghasilkan data baru secara otomatis, beberapa hasil augmentasi menunjukkan perubahan kata yang kurang sesuai dengan konteks kalimat. Sebagai contoh, frasa *"menyesuaikan diri"* diubah menjadi *"bela*

diri" sehingga makna kalimat mengalami pergeseran. Kasus serupa juga terlihat pada penggantian kata *"teknologi"* menjadi *"diri"* yang menghasilkan kalimat kurang natural dan berpotensi mengubah informasi yang terkandung pada teks asli. Sebaliknya, metode *Back Translation* menghasilkan variasi kalimat melalui proses penerjemahan ke bahasa perantara dan penerjemahan kembali ke bahasa asal. Pendekatan ini cenderung mempertahankan makna utama kalimat meskipun terjadi perubahan susunan kata atau bentuk ekspresi. Hasil augmentasi yang dihasilkan terlihat lebih alami dan mendekati variasi bahasa yang umum digunakan dalam jawaban peserta didik. Karakteristik tersebut menunjukkan bahwa *Back Translation* mampu menghasilkan data dengan kualitas semantik yang lebih baik dibandingkan augmentasi berbasis substitusi kata.

Temuan ini mengindikasikan bahwa kualitas variasi kalimat yang dihasilkan oleh metode augmentasi dapat menjadi faktor penting yang memengaruhi performa model klasifikasi. Meskipun kedua metode mampu meningkatkan jumlah data dan menghasilkan distribusi kelas yang seimbang, *Back Translation* memiliki potensi lebih besar dalam mempertahankan makna asli kalimat sehingga dapat membantu model mempelajari representasi bahasa yang lebih baik.

3.3 Hasil Pengujian *Non K-Fold*

Pengujian *Non K-Fold* dilakukan menggunakan data pengujian (*testing set*) yang telah ditentukan pada dataset UKARA. Pada skenario ini, model dilatih menggunakan data hasil augmentasi dan dievaluasi menggunakan data pengujian yang tidak melalui proses augmentasi. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian *Non K-Fold*

Dataset	**Metode**	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1-Score (%)***
A	EDA Sinonim IndoBERT	82,83	83,67	82,83	83.12
A	*Back Translation*	89,00	89,00	89,00	89,00
B	EDA Sinonim IndoBERT	70,16	71,34	70.16	68.10
B	*Back Translation*	71,43	71,00	71,00	71,00

Berdasarkan Tabel 5, metode *Back Translation* menghasilkan performa yang lebih baik dibandingkan EDA Sinonim IndoBERT pada kedua dataset yang digunakan. Pada Dataset A, *Back Translation* memperoleh akurasi sebesar 89,00%, meningkat 6,17% dibandingkan EDA Sinonim IndoBERT yang memperoleh akurasi sebesar 82,83%. Peningkatan yang serupa juga terlihat pada metrik *precision*, *recall*, dan *F1-score,* yang seluruhnya mencapai nilai 89,00%. Pada Dataset B, *Back Translation* juga menunjukkan performa yang lebih baik dengan akurasi sebesar 71,43%, sedangkan EDA Sinonim IndoBERT memperoleh akurasi sebesar 70,16%. Meskipun peningkatannya relatif kecil, yaitu sebesar 1,27%, *Back Translation* tetap menghasilkan nilai F1-score yang lebih tinggi, yaitu 71,00% dibandingkan 68,10% pada EDA Sinonim IndoBERT.

Hasil tersebut menunjukkan bahwa variasi kalimat yang dihasilkan melalui *Back Translation* mampu memberikan representasi data yang lebih baik dibandingkan augmentasi berbasis substitusi kata menggunakan EDA Sinonim IndoBERT. Meskipun kedua metode menghasilkan jumlah data dan distribusi kelas yang relatif seimbang, *Back Translation* menghasilkan variasi pada tingkat kalimat dan struktur sintaksis sehingga model dapat mempelajari pola bahasa yang lebih beragam. Temuan ini mengindikasikan bahwa kualitas variasi data hasil augmentasi memiliki pengaruh yang lebih besar terhadap performa model dibandingkan sekadar peningkatan jumlah data pelatihan.

3.4 Hasil Pengujian *3-Fold Cross Validation*

Untuk memperoleh evaluasi yang lebih stabil, model diuji menggunakan metode *3-Fold Cross Validation.* Pada skenario ini, data pelatihan (*training set*) dan data validasi (*validation set*) digabungkan terlebih dahulu, kemudian dibagi menjadi tiga subset yang digunakan secara bergantian

sebagai data pelatihan dan data pengujian. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian ditunjukkan pada Tabel 6.

Tabel 6. Hasil Pengujian *3-Fold Cross Validation*

Dataset	Metode	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
A	EDA Sinonim IndoBERT	82,83	82,98	82,84	82,90
A	Back Translation	86,96	87,12	86,96	86,94
B	EDA Sinonim IndoBERT	72,13	72,12	72,13	71,83
B	Back Translation	71,43	73,17	71,43	70,88

Berdasarkan Tabel 6, metode *Back Translation* menunjukkan performa yang lebih baik pada Dataset A dengan akurasi sebesar 86,96%, lebih tinggi dibandingkan EDA Sinonim IndoBERT yang memperoleh akurasi sebesar 82,83%. Peningkatan tersebut juga terlihat pada metrik *precision, recall*, dan *F1-score*, yang menunjukkan bahwa variasi data yang dihasilkan melalui *Back Translation* mampu meningkatkan kemampuan model dalam mengenali pola jawaban pada Dataset A. Berbeda dengan Dataset A, pada Dataset B metode EDA Sinonim IndoBERT memperoleh performa yang sedikit lebih baik dengan akurasi sebesar 72,13%, sedangkan *Back Translation* memperoleh akurasi sebesar 71,43%. Selisih performa yang relatif kecil menunjukkan bahwa kedua metode memiliki kemampuan yang hampir setara pada dataset ini. Hasil tersebut mengindikasikan bahwa efektivitas metode augmentasi tidak hanya dipengaruhi oleh teknik augmentasi yang digunakan, tetapi juga oleh karakteristik data yang menjadi objek pembelajaran model.

Secara umum, hasil pengujian *3-Fold Cross Validation* menunjukkan bahwa kedua metode augmentasi mampu meningkatkan keragaman data dan menghasilkan performa yang kompetitif. Namun, *Back Translation* cenderung memberikan keuntungan yang lebih besar pada Dataset A, sedangkan EDA Sinonim IndoBERT menunjukkan performa yang sedikit lebih baik pada Dataset B. Temuan ini menunjukkan bahwa tidak terdapat satu metode augmentasi yang secara konsisten unggul pada seluruh kondisi dataset, sehingga karakteristik data menjadi faktor penting dalam menentukan efektivitas metode augmentasi yang digunakan.

3.5 Pembahasan

Hasil penelitian menunjukkan bahwa metode augmentasi memberikan pengaruh yang berbeda terhadap performa model BiLSTM pada masing-masing dataset. Pada skenario *Non K-Fold*, *Back Translation* menghasilkan performa yang lebih baik dibandingkan EDA Sinonim IndoBERT pada Dataset A maupun Dataset B. Peningkatan terbesar terjadi pada Dataset A dengan kenaikan akurasi dari 82,83% menjadi 89,00%, sedangkan pada Dataset B peningkatan akurasi yang diperoleh relatif kecil, yaitu dari 70,16% menjadi 71,43%. Hasil yang berbeda ditemukan pada pengujian menggunakan *3-Fold Cross Validation. Back Translation* tetap menghasilkan performa terbaik pada Dataset A dengan akurasi sebesar 86,96%, lebih tinggi dibandingkan EDA Sinonim IndoBERT yang memperoleh akurasi sebesar 82,83%. Namun, pada Dataset B EDA Sinonim IndoBERT memperoleh akurasi yang sedikit lebih tinggi, yaitu 72,13%, dibandingkan *Back Translation* sebesar 71,43%. Meskipun demikian, selisih performa yang diperoleh relatif kecil sehingga kedua metode dapat dianggap memiliki kemampuan yang hampir setara pada Dataset B.

Perbedaan hasil tersebut menunjukkan bahwa efektivitas metode augmentasi sangat dipengaruhi oleh karakteristik dataset yang digunakan. Pada Dataset A yang memiliki tingkat ketidakseimbangan kelas lebih tinggi sebelum augmentasi, *Back Translation* mampu menghasilkan variasi kalimat yang lebih beragam melalui perubahan struktur sintaksis tanpa mengubah makna utama kalimat. Variasi tersebut membantu model mempelajari representasi bahasa yang lebih luas sehingga meningkatkan kemampuan generalisasi pada proses klasifikasi. Temuan ini sejalan dengan penelitian Sennrich et al.

(2016) yang menunjukkan bahwa *Back Translation* mampu meningkatkan kualitas data pelatihan melalui pembentukan variasi kalimat yang lebih alami dibandingkan teknik augmentasi berbasis manipulasi kata. Di sisi lain, EDA Sinonim IndoBERT menghasilkan variasi data melalui penggantian kata menggunakan sinonim yang relevan secara kontekstual. Pendekatan ini terbukti tetap efektif pada Dataset B, terutama ketika jumlah data yang tersedia relatif lebih besar dan distribusi kelas lebih seimbang. Hasil tersebut mengindikasikan bahwa augmentasi pada tingkat kata masih mampu memberikan manfaat pada kondisi tertentu, terutama ketika variasi jawaban tidak memerlukan perubahan struktur kalimat yang kompleks.

Menariknya, hasil penelitian juga menunjukkan bahwa jumlah data hasil augmentasi tidak selalu berbanding lurus dengan performa model. Pada Dataset A, EDA Sinonim IndoBERT menghasilkan 1.218 data, sedangkan *Back Translation* menghasilkan 1.220 data dengan performa yang lebih tinggi. Temuan ini menunjukkan bahwa kualitas variasi data yang dihasilkan memiliki peran yang lebih penting dibandingkan sekadar peningkatan jumlah data. Dengan kata lain, kemampuan metode augmentasi dalam menghasilkan variasi kalimat yang tetap mempertahankan makna asli menjadi faktor utama yang memengaruhi keberhasilan model klasifikasi.

Secara keseluruhan, hasil penelitian menunjukkan bahwa *Back Translation* merupakan alternatif yang efektif untuk augmentasi data pada sistem *Automatic Short Answer Scoring* Bahasa Indonesia. Meskipun tidak selalu menghasilkan performa terbaik pada seluruh dataset, metode ini mampu memberikan peningkatan performa yang lebih konsisten dibandingkan EDA Sinonim IndoBERT, khususnya pada dataset dengan tingkat ketidakseimbangan kelas yang lebih tinggi.

4. KESIMPULAN DAN SARAN

Penelitian ini telah menganalisis efektivitas metode *Back Translation* sebagai teknik augmentasi dataset pada sistem *Automatic Short Answer Scoring* (AES) Bahasa Indonesia menggunakan dataset UKARA. Hasil penelitian menunjukkan bahwa *Back Translation* mampu memberikan performa yang lebih baik dibandingkan EDA Sinonim IndoBERT pada sebagian besar skenario pengujian. Pada pengujian *Non K-Fold*, *Back Translation* memperoleh akurasi sebesar 89,00% pada Dataset A dan 71,43% pada Dataset B, lebih tinggi dibandingkan EDA Sinonim IndoBERT yang memperoleh akurasi sebesar 82,83% dan 70,16%. Pada pengujian 3-*Fold Cross Validation, Back Translation* juga menunjukkan performa terbaik pada Dataset A dengan akurasi 86,96%, sedangkan pada Dataset B EDA Sinonim IndoBERT memperoleh hasil yang sedikit lebih tinggi dengan akurasi 72,13%.

Hasil penelitian menunjukkan bahwa peningkatan performa model tidak hanya dipengaruhi oleh jumlah data hasil augmentasi, tetapi juga oleh kualitas variasi data yang dihasilkan. *Back Translation* mampu menghasilkan variasi kalimat yang lebih alami melalui perubahan struktur sintaksis tanpa mengubah makna utama teks, sehingga memberikan peningkatan performa yang lebih konsisten dibandingkan augmentasi berbasis substitusi kata. Dengan demikian, *Back Translation* dapat menjadi alternatif yang efektif untuk meningkatkan kualitas dataset pada sistem penilaian otomatis tes uraian singkat Bahasa Indonesia.

Penelitian selanjutnya dapat mengkaji penggunaan bahasa perantara yang berbeda dalam proses *Back Translation* atau mengombinasikannya dengan metode augmentasi lainnya untuk menghasilkan variasi data yang lebih beragam. Selain itu, pengujian pada dataset yang lebih besar dan memiliki karakteristik yang berbeda perlu dilakukan untuk memperoleh pemahaman yang lebih komprehensif mengenai efektivitas metode augmentasi pada sistem *Automatic Short Answer Scoring* Bahasa Indonesia.

REFERENSI

Ariyus, D., Manongga, D., & Sembiring, I. (2024). Enhancing Sentiment Analysis of Indonesian Tourism Video Content Commentary on TikTok: A FastText and Bi-LSTM Approach.

Engineering, Technology & Applied Science Research, 14(6), 18020-18028. https://doi.org/10.48084/etasr.8859

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115.

Fadilah, N., & Priyanta, S. (2022). Automatic Essay Scoring Using Data Augmentation in Bahasa Indonesia. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 16(4), 401-410.

Herwanto, G. B., Sari, Y., Prastowo, B. N., Bustoni, I. A., & Hidayatulloh, I. (2018). UKARA: A fast and simple automatic short answer scoring system for Bahasa Indonesia. *ICEAP 2019*, 2, 48-53.

Nurhaliza, S., & Al Fatta, H. (2024). Analisis Sentimen Ulasan Aplikasi KAI Access Menggunakan Metode Naive Bayes Classifier dan Support Vector Machine dengan Teknik Preprocessing. *Jurnal Teknoinfo*, 18(1), 145-156.

Purwoko, R. A., Utami, E., & Luthfi, E. T. (2023). Klasifikasi Teks Aduan Publik Menggunakan Kombinasi TF-IDF dan Bidirectional Long Short-Term Memory (BiLSTM). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(4), 812-820.

Rahma, I. A., & Suadaa, L. H. (2023). Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 10(6), 1329-1340.

Sani, D. A., & Sarwani, M. Z. (2022). Koreksi Jawaban Esai Berdasarkan Persamaan Makna Menggunakan Fasttext dan Algoritma Backpropagation. Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), 11(2), 92–111.

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Siahaan, M., Sitanggang, F., & Girsang, A. S. (2023). Metode Komparatif Teknik Augmentasi Data pada Klasifikasi Teks Sentimen Berbahasa Indonesia Menggunakan Dataset Terbatas. *Jurnal Media Informatika Budidarma*, 7(2), 743-752.

Tanaka, E. A., Anderies, A., & Chowanda, A. (2024). Evaluating Back Translation and Misspelling Correction Utilization on Indonesian AES. In *2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 1-6). IEEE.

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv preprint arXiv:1901.11196*.