Fine-Grained Sentiment Analysis Menggunakan *Text Augmentation* untuk Ulasan Restoran

Fine-Grained Sentiment Analysis Using Text Augmentation for Restaurant Reviews

Audrey Sasqhia Wijaya[1], **Amanda Illona Farrel**[2]
[1,2]*Departemen Teknik Informatika, Institut Teknologi Sepuluh Nopember*
E-mail : 5025221055@student.its.ac.id[1], 5025221056@student.its.ac.id@site.com[2]
Received 2 December 2025; Revised 4 December 2025; Accepted 10 December 2025

Abstrak - Penelitian ini mengkaji klasifikasi rating restoran berbasis ulasan teks dari Google Maps dengan target lima kelas. Tantangan utama pada tugas ini adalah batas antar kelas yang tipis, sifat label yang bersifat ordinal, serta kemungkinan ketidakseimbangan distribusi kelas. Dataset yang digunakan berasal dari Kaggle (Google Maps Restaurant Reviews) dan diproses melalui normalisasi teks dan lemmatization. Evaluasi dilakukan menggunakan pembagian data 80:20 dengan stratified split dan diukur menggunakan metrik Accuracy, F1-macro, dan F1-weighted. Eksperimen dibangun dalam tiga skenario: (1) baseline TF-IDF dengan model Linear SVM, Naive Bayes, dan Logistic Regression; (2) baseline Transformer menggunakan BERT dan DistilBERT; serta (3) penerapan augmentasi pada data latih menggunakan EDA, Modified EDA, BERT augmentation, dan backtranslation. Hasil menunjukkan bahwa model Transformer secara konsisten mengungguli baseline TF-IDF. Baseline TF-IDF, Linear SVM menghasilkan kinerja terbaik dengan F1-macro 0,3758. Sementara itu, baseline Transformer menghasilkan performa lebih tinggi, dengan BERT mencapai F1-macro 0,4203. Peningkatan paling baik diperoleh pada skenario augmentasi, di mana BERT dengan EDA menghasilkan F1-macro tertinggi sebesar 0,4761 dan akurasi 0,5636. Temuan ini menegaskan bahwa pemahaman konteks kalimat berperan penting pada klasifikasi rating 1–5 dan bahwa augmentasi yang menjaga naturalitas teks lebih efektif, terutama ketika dikombinasikan dengan model berbasis Transformer.

Kata kunci - *Sentiment Analysis*, *Transformer*, *Ulasan, TF-IDF, BERT, Augmentasi*

Abstract - *This study investigates five-class restaurant rating classification using textual reviews from Google Maps. The task is challenging due to subtle boundaries between adjacent classes, the ordinal nature of ratings, and potential class imbalance. The dataset was obtained from Kaggle (Google Maps Restaurant Reviews) and processed through text normalization and lemmatization. Model performance was evaluated using an 80:20 stratified train-test split and measured with Accuracy, Macro-F1, and Weighted-F1. Three experimental scenarios were designed: (1) TF-IDF baselines using Linear SVM, Naive Bayes, and Logistic Regression; (2) Transformer baselines using BERT and DistilBERT; and (3) data augmentation on the training set using EDA, Modified EDA, BERT-based augmentation, and backtranslation. Results show that Transformer-based models consistently outperform TF-IDF baselines. In the TF-IDF scenario, Linear SVM achieved the best baseline performance with a Macro-F1 of 0.3758. The Transformer baseline improved overall performance, with BERT reaching a Macro-F1 of 0.4203. The most effective improvement was observed in the augmentation scenario, where BERT combined with EDA achieved the highest Macro-F1 of 0.4761 and an accuracy of 0.5636. DistilBERT with backtranslation also showed competitive performance with a Macro-F1 of 0.4462. These findings emphasize the importance of contextual modeling for fine-grained 1–5 star classification and suggest that augmentation strategies preserving sentence naturalness are particularly beneficial when paired with Transformer architectures.*

Keywords: *Sentiment Ana*lysis, *Transformer*, *Review, TF-IDF, BERT, Augmentation*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

1. PENDAHULUAN

Ulasan pelanggan pada platform daring merupakan data tekstual yang memuat informasi mengenai kualitas layanan, harga, dan produk. Analisis sentimen biner memiliki keterbatasan dalam merepresentasikan tingkat kepuasan pelanggan pada skala spesifik. *Fine-Grained Sentiment Analysis* (FGSA) dengan klasifikasi skala 1 hingga 5 memberikan granularitas informasi bagi strategi bisnis [1]. Kendala utama FGSA pada data publik adalah distribusi kelas tidak seimbang. Dominasi kelas mayoritas dibandingkan kelas minoritas menyebabkan bias prediksi pada model klasifikasi dan kegagalan generalisasi pola data [2].

Penanganan ketidakseimbangan data umumnya menggunakan metode *oversampling* pada level fitur. Metode ini tidak menghasilkan variasi linguistik pada data teks. *Text Data Augmentation* (TDA) menjadi pendekatan alternatif untuk meningkatkan kuantitas data latih secara sintetis tanpa pengumpulan data manual. Metode TDA meliputi *Easy Data Augmentation* (EDA) dengan operasi penyisipan dan pertukaran kata, serta *Backtranslation* untuk menghasilkan parafrasa [3]. Penerapan augmentasi data bertujuan meningkatkan performa model pada dataset dengan sumber daya terbatas (*low resource*) [4].

Kinerja analisis sentimen bergantung pada arsitektur model klasifikasi. Model klasik seperti *Support Vector Machine* (SVM), *Naive Bayes*, dan *Logistic Regression* memiliki efisiensi komputasi tinggi. Model berbasis *Transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*) dan DistilBERT menggunakan mekanisme *attention* untuk pemahaman konteks [5]. Penelitian [6] menerapkan augmentasi berbasis IndoBERT untuk klasifikasi teks dan menghasilkan data sintetik yang mempertahankan makna semantik. Sementara itu, penelitian lain [7] menganalisis augmentasi berbasis token pada bahasa India dan mencatat peningkatan kinerja pada model Bi-LSTM, namun dampak minimal pada model *pre-trained* seperti IndicBERT. Selanjutnya, penelitian [8] mengusulkan *Modified EDA* dengan integrasi *Part-of-Speech* (POS) *tagging* pada ulasan restoran untuk menjaga konsistensi label sentimen.

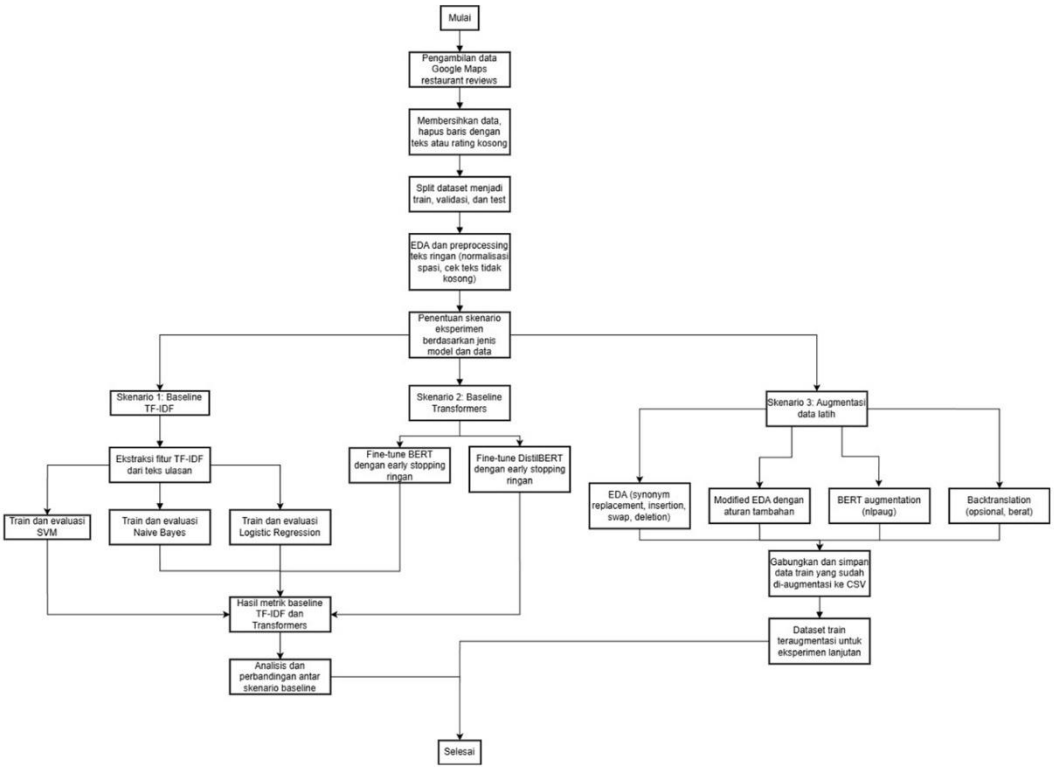
Penelitian terdahulu membuktikan efektivitas augmentasi data, namun komparasi spesifik pada kasus *fine-grained sentiment analysis* berskala 1 hingga 5 masih terbatas. Studi sebelumnya belum membandingkan efektivitas antara augmentasi acak (*Original EDA*), augmentasi berbasis aturan (*Modified EDA*), augmentasi berbasis konteks (*BERT Augmentation*), dan *backtranslation* secara signifikan terhadap kinerja model klasik dan model *Transformer*. Ketiadaan analisis komparatif ini menjadi dasar pelaksanaan penelitian.

Penelitian ini bertujuan meningkatkan akurasi *Fine-Grained Sentiment Analysis* pada dataset ulasan restoran tidak seimbang menggunakan pendekatan *Text Augmentation*. Kebaruan penelitian terletak pada analisis komparatif dampak penerapan *Original EDA*, *Modified EDA*, *BERT Augmentation*, dan *backtranslation* terhadap model klasik (*Support Vector Machine, Naive Bayes*, *Logistic Regression*) dan model berbasis *Transformer* (BERT, DistilBERT). Hasil penelitian memberikan evaluasi empiris mengenai teknik augmentasi dalam meningkatkan performa klasifikasi pada kelas minoritas.

2. METODE PENELITIAN

Metodologi penelitian ini terdiri dari beberapa tahapan yang tersusun secara sistematis, mulai dari proses pengumpulan data, pengolahan, hingga analisis hasil. Setiap tahapan dirancang untuk memastikan bahwa penelitian dilakukan secara terstruktur, terukur, dan dapat dipertanggungjawabkan. Dengan penyusunan langkah-langkah tersebut, penelitian ini diharapkan mampu memberikan gambaran menyeluruh mengenai alur kerja yang dilakukan, sekaligus memperjelas bagaimana data diolah menjadi temuan yang bermakna. Rangkaian proses metodologis ini ditampilkan secara visual pada Gambar 1 untuk memudahkan pemahaman terhadap keseluruhan proses penelitian.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :



Gambar 1. Diagram alir Proses

2.1 Dataset

Penelitian ini menggunakan dataset dari Kaggle dengan nama *Google Maps Restaurant Reviews*, yang berisi sekitar 1.100 ulasan restoran dalam bahasa Inggris. Setiap entri mencakup teks ulasan, rating bintang, identitas restoran, serta foto yang diunggah oleh pengguna. Dataset ini juga menyediakan kategori tambahan yang berkaitan dengan konten visual pada foto ulasan. Meskipun dataset terdiri dari beberapa berkas, penelitian ini hanya menggunakan berkas yang memuat data ulasan pelanggan.

Tabel 1. Deskripsi Dataset

Fitur	Deskripsi data
business_name	Identitas restoran yang diulas oleh pengguna
author_name	Nama pengguna yang memberikan ulasan
text	Pernyataan atau komentar pengguna mengenai pengalaman mereka
photo	Gambar yang diunggah bersama ulasan (jika tersedia)
rating	Nilai penilaian dalam skala 1–5 yang diberikan oleh pengguna
rating_category	Kategori ulasan berdasarkan aspek pada foto (misalnya rasa, menu, suasana dalam, suasana luar)

2.2 Data Preprocessing

Tahap pra-pemrosesan data (data preprocessing) bertujuan untuk mentransformasi data mentah menjadi format terstruktur dan bersih agar optimal saat diproses oleh algoritma machine learning. Proses ini diklasifikasikan menjadi dua tahapan utama, yaitu pembersihan data (data cleaning) dan pra-pemrosesan teks.

2.2.1 Pembersihan Data

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

Sebelum memasuki tahap pemrosesan teks, dilakukan filtrasi awal untuk menjamin kualitas data. Langkah pertama adalah penanganan nilai yang hilang (missing values) dengan mengeliminasi baris data yang mengandung nilai null pada kolom text dan rating. Selanjutnya, dilakukan validasi pada kolom rating untuk memastikan tipe data berupa integer dan berada dalam rentang valid (skala 1 sampai 5). Data yang berada di luar rentang tersebut atau tidak memiliki teks ulasan yang valid akan dianulir dari dataset.

2.2.2 Pemrosesan Teks

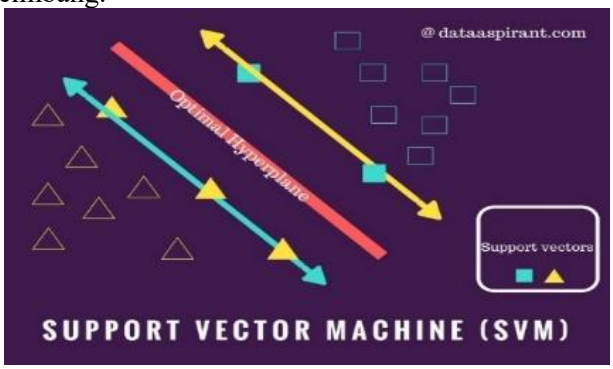
Pada bagian ini, kolom teks diproses lebih lanjut menggunakan library *NLTK* untuk menstandarkan data dan mereduksi variasi fitur kata. Proses ini diawali dengan tahap *case folding* dan *noise removal*, di mana seluruh huruf dikonversi menjadi huruf kecil (*lowercase*) serta karakter non-alfabet seperti angka dan tanda baca dihapus menggunakan *regular expression*, sehingga hanya menyisakan karakter huruf dengan spasi tunggal [9]

Teks yang telah dibersihkan kemudian dipecah menjadi unit kata melalui proses *tokenization*. Untuk meningkatkan akurasi pada tahap selanjutnya, dilakukan *Part-of-Speech (POS) tagging* untuk memberikan label kelas kata pada setiap token, seperti *noun*, *verb*, *adjective*, atau *adverb*, yang berfungsi membedakan konteks penggunaan kata ada [10].

Tahapan ini diakhiri dengan proses *lemmatization* menggunakan *WordNet Lemmatizer*, yang mengubah kata menjadi bentuk dasarnya (*lemma*) dengan mempertimbangkan *POS tag* yang telah diidentifikasi. Pendekatan ini lebih akurat dibandingkan *stemming*; sebagai contoh, kata "better" akan dikembalikan menjadi "good" dan "running" menjadi "run" [11].

2.3 Model Development

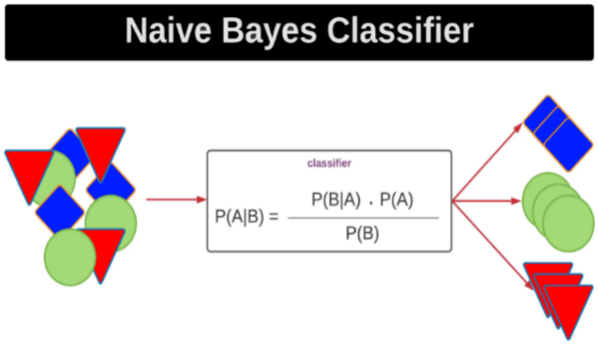
Support Vector Machine (SVM) merupakan algoritma pembelajaran yang memisahkan kelas melalui penentuan *hyperplane* optimal dalam ruang fitur berdimensi tinggi. SVM linier menangani data dimensi besar dan jarang (*sparse*) seperti representasi vektor TF-IDF. Algoritma ini meminimalkan risiko kesalahan dan memaksimalkan margin geometris antar kelas. Mekanisme tersebut mencegah *overfitting* pada dataset dengan jumlah sampel terbatas atau distribusi yang tidak seimbang.



Gambar 2. Ilustrasi Model Klasik SVM (Saxena, 2017)

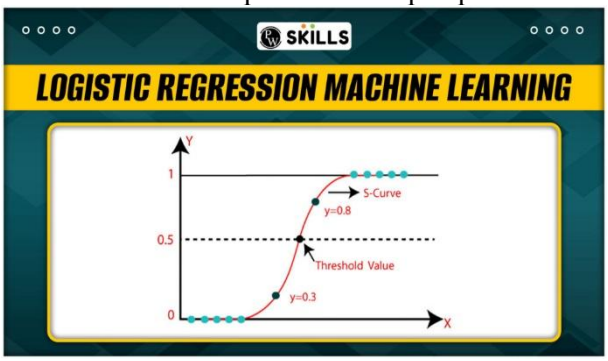
Naive Bayes adalah metode klasifikasi probabilistik berbasis Teorema Bayes dengan asumsi independensi antar fitur. Meskipun asumsi independensi antar kata jarang terpenuhi secara penuh dalam teks, algoritma ini berfungsi sebagai *baseline* klasifikasi karena efisiensi komputasi yang tinggi. *Naive Bayes* menghitung probabilitas posterior setiap kelas berdasarkan frekuensi kemunculan kata. Pendekatan ini meminimalkan penggunaan memori pada pemrosesan dataset berskala besar.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :



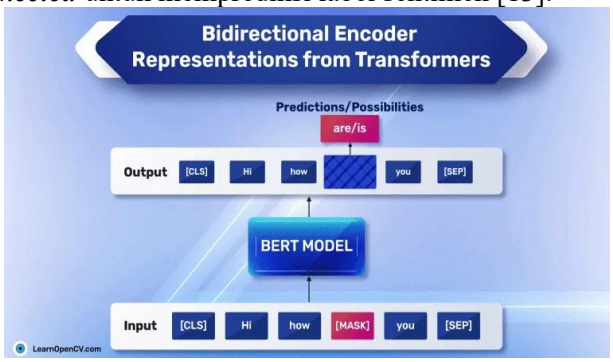
Gambar 3. Ilustrasi Model Klasik Naïve Bayes (Elzeiny, 2023)

Logistic Regression memprediksi probabilitas kategori kelas menggunakan fungsi logistik atau *sigmoid*. Model ini memetakan nilai prediksi ke dalam rentang 0 hingga 1 sebagai representasi probabilitas keanggotaan kelas. Pada klasifikasi *fine-grained* multi-kelas, algoritma ini menerapkan skema *One-vs-Rest* (OvR) atau *Multinomial Logistic Regression*. Nilai bobot fitur dalam model menunjukkan kontribusi setiap kata terhadap keputusan klasifikasi sentimen [12].



Gambar 4. Ilustrasi Model Klasik Logistic Regression (Sumber: https://blog.pwskills.com/wp-content/uploads/2024/06/Logistic-Regression-Machine-Learning-1024x576.jpg)

Bidirectional Encoder Representations from Transformers (BERT) menggunakan arsitektur *Transformer* dengan mekanisme *self-attention* untuk memproses konteks kata secara dua arah (*bidirectional*). Model ini melalui proses *pre-training* pada teks menggunakan *Masked Language Modeling* (MLM) untuk menangkap struktur sintaksis dan semantik. Implementasi BERT pada klasifikasi melibatkan tahapan *fine-tuning*, di mana representasi vektor token klasifikasi menjadi input lapisan *fully connected* untuk memprediksi label sentimen [13].

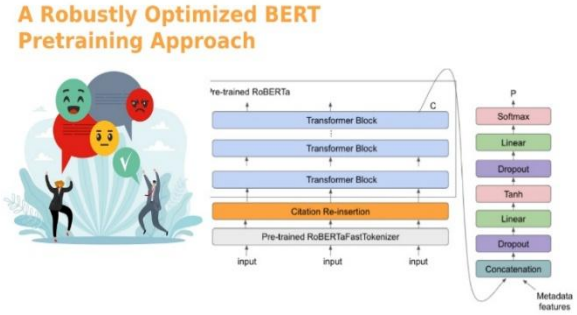


Gambar 5. Ilustrasi Model Transformer BERT (Rath, 2023)

DistilBERT merupakan varian model hasil teknik *knowledge distillation* dari arsitektur BERT dasar. Model ini mempertahankan kapabilitas pemahaman bahasa BERT dengan reduksi

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

parameter sebesar 40% dan peningkatan kecepatan inferensi sebesar 60%. DistilBERT melatih modelnya untuk meniru distribusi probabilitas keluaran dari BERT. Mekanisme ini menghasilkan efisiensi komputasi tinggi untuk penerapan pada perangkat dengan sumber daya terbatas tanpa penurunan akurasi yang signifikan [14].



Gambar 6. Ilustrasi Model Transformer DistilBERT (Naveed, 2023)

2.4 Text Data Augmentation (TDA)

Text Data Augmentation (TDA) diterapkan sebagai strategi untuk mengatasi kelangkaan dan ketidakseimbangan distribusi data latih dengan menghasilkan variasi data sintetik yang mempertahankan makna semantik asli [15]. Penelitian ini mengintegrasikan empat teknik augmentasi spesifik guna meningkatkan generalisasi model. Pertama, *Original Easy Data Augmentation* (EDA) menerapkan operasi stokastik seperti penggantian sinonim dan penyisipan acak untuk meningkatkan variabilitas data secara efisien [16]. Kedua, *Backtranslation* digunakan untuk menghasilkan parafrasa melalui proses penerjemahan dua arah yang memperkaya keragaman struktur sintaksis tanpa mengubah inti informasi [17]. Ketiga, untuk memitigasi risiko perubahan polaritas sentimen akibat operasi acak, diterapkan *Modified EDA* yang mengintegrasikan aturan *Part-of-Speech* (POS) *tagging* guna memproteksi kata kunci bermuatan sentimen dari modifikasi. Terakhir, *BERT Augmentation* memanfaatkan representasi kontekstual dari model *Transformer* untuk memprediksi dan mensubstitusi token yang ditutupi (*masked*) dengan kata yang paling relevan secara semantik, menghasilkan variasi kalimat yang lebih koheren.

2.5 Evaluasi Model

Evaluasi model merupakan proses untuk mengukur performa model yang dihasilkan pada tahap pelatihan dan pengujian [18]. Tahap ini bertujuan untuk mengetahui sejauh mana model mampu menghasilkan prediksi yang akurat terhadap data uji. Pada penelitian ini digunakan dua metrik utama, yaitu *Accuracy* dan *F1-Score* (*Macro* dan *Weighted*).

Accuracy mengukur proporsi prediksi yang benar terhadap seluruh prediksi yang dilakukan. Metrik ini umum digunakan ketika setiap label dianggap memiliki tingkat kepentingan yang sama serta ketika prediksi *True Positive* (*TP*) dan *True Negative* (*TN*) memiliki peran penting. *Accuracy* dihitung menggunakan (1),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

di mana *TP* adalah jumlah data positif yang diprediksi benar sebagai positif, *TN* adalah jumlah data negatif yang diprediksi benar sebagai negatif, *FP* adalah jumlah data negatif yang salah diprediksi sebagai positif, dan *FN* adalah jumlah data positif yang salah diprediksi sebagai negatif.

Sementara itu, *F1-Score* merupakan rata-rata harmonik antara *precision* dan *recall*, yang memberikan penilaian lebih seimbang terutama pada dataset dengan distribusi kelas tidak merata. Perhitungan *F1-Score* diberikan oleh (2),

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

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

di mana *precision* menunjukkan proporsi prediksi positif yang benar-benar positif, sedangkan *recall* menunjukkan kemampuan model dalam menemukan seluruh data positif yang sebenarnya ada.

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Sub Bab ini memaparkan hasil evaluasi kinerja model klasifikasi *fine-grained sentiment analysis* (rating 1-5) pada dataset ulasan restoran. Evaluasi dilakukan menggunakan metrik *Accuracy*, *F1-Macro*, dan *F1-Weighted*. *F1-Macro* menjadi metrik fokus utama mengingat kondisi dataset yang tidak seimbang (*imbalanced*), di mana kemampuan model dalam mengenali kelas minoritas menjadi prioritas. Pengujian dibagi menjadi tiga skenario: (1) *Baseline* Model Klasik (TF-IDF), (2) *Baseline* Model Transformer, dan (3) Eksperimen Augmentasi Data

3.1.1 Skenario 1 Baseline TF-IDF

Skenario ini mengevaluasi kinerja tiga model pembelajaran mesin klasik, yaitu *Linear Support Vector Machine*, *Naive Bayes*, dan *Logistic Regression* menggunakan ekstraksi fitur TF-IDF tanpa augmentasi data. Ringkasan hasil disajikan pada tabel berikut.

Tabel 2. Rangkuman Model Skenario 1 Baseline TF-IDF

Model	*Accuracy*	*F1 Macro*	*F1 Weighted*
Linear SVM (S1 baseline)	**0.5000**	**0.3758**	**0.4734**
Logistic Regression (S1 baseline)	0.4773	0.2396	0.4025
Naive Bayes (S1 baseline)	0.4364	0.1479	0.2912

Hasil menunjukkan bahwa *Linear Support Vector Machine* memberikan kinerja terbaik di antara model klasik, dengan akurasi sebesar 50% dan *F1-Macro* 0.3758. Sementara itu, *Naive Bayes* mencatat performa terendah dengan *F1-Macro* 0.1479, yang menunjukkan keterbatasannya dalam menangani distribusi kelas yang tidak seimbang pada dataset ini.

3.1.2 Skenario 2 Baseline Transformer

Skenario ini menerapkan model *Deep Learning* berbasis Transformer, yaitu BERT dan DistilBERT. Proses pelatihan dijalankan menggunakan algoritma optimasi AdamW dengan *learning rate* rendah untuk menjaga stabilitas bobot *pre-trained*. Mekanisme *Early Stopping* diterapkan untuk mengefisienkan waktu komputasi. Rincian konfigurasi *hyperparameter* pelatihan disajikan pada tabel berikut.

Tabel 3. Konfigurasi Model Transformer

Parameter	Keterangan
Model Pre-trained	bert-base-uncased, distilbert-base-uncased
Optimizer	AdamW
Learning Rate	2×10−5
Batch Size	8
Max Sequence Length	256 token
Max Epochs	100 (dengan *Early Stopping*)
Early Stopping Patience	1 epoch
Loss Function	Cross-Entropy Loss
Best Model Metric	Macro F1-Score

Berdasarkan konfigurasi tersebut, model menggunakan pendekatan *fine-tuning* pada data asli dan hasil pengujian disajikan pada tabel berikut.

Tabel 4. Rangkuman Model Skenario 2 Baseline Transformer

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

Model	*Accuracy*	*F1 Macro*	*F1 Weighted*	*Eval Loss*	*Epoch*
bert-base-uncased	5.364	**4.203**	**5.338**	1.3769	6.0
distilbert-base-uncased	**5.409**	3.845	5.233	**1.1194**	5.0

Secara umum, model *Transformer* mengungguli model klasik. *DistilBERT* unggul tipis dalam akurasi (54.09%) dan efisiensi komputasi (*loss* terendah), namun *BERT* menunjukkan kemampuan generalisasi antar-kelas yang lebih baik dengan skor *F1-Macro* tertinggi (0.4203).

3.1.3 Skenario 3

Skenario ini menerapkan augmentasi untuk meningkatkan variasi data latih dan mengurangi dampak ketidakseimbangan kelas. Empat strategi augmentasi yang digunakan adalah *EDA*, *Modified EDA*, *BERT Augmentation*, dan *Backtranslation*. Augmentasi kemudian dievaluasi pada *Linear Support Vector Machine*, *Naive Bayes*, *Logistic Regression*, *BERT,* dan *DistilBERT*.

Tabel 5. Rangkuman Model BERT Skenario 3

Augmentation	*Accuracy*	*F1 Macro*	*F1 Weighted*	*Eval Loss*	*Epoch*
EDA	0.5636	**0.4761**	0.5528	1.3251	4.0
Backtranslation	0.5636	0.4009	0.5520	1.2118	3.0
Modified EDA	0.5409	0.3751	0.5025	1.2145	3.0
BERT Aug	0.4864	0.3472	0.4421	1.2363	2.0

Tabel 6. Rangkuman Model DistilBERT Skenario 3

Augmentation	*Accuracy*	*F1 Macro*	*F1 Weighted*	*Eval Loss*	*Epoch*
Backtranslation	0.5545	**0.4462**	0.5551	1.1702	3.0
EDA	0.5318	0.4458	0.5298	1.1342	3.0
BERT Aug	0.5091	0.4190	0.5124	1.2668	4.0
Modified EDA	0.5182	0.4150	0.5184	1.1827	3.0

Tabel 7. Rangkuman Model Linear SVM Skenario 3

Augmentation	*Accuracy*	*F1 Macro*	*F1 Weighted*
Modified EDA	0.4682	**0.3809**	0.4601
EDA	0.4727	0.3751	0.4649
Backtranslation	0.4591	0.3739	0.4539
BERT Aug	0.4409	0.3727	0.4377

Tabel 8. Rangkuman Model Naïve Bayes Skenario 3

Augmentation	*Accuracy*	*F1 Macro*	*F1 Weighted*
Modified EDA	0.4636	**0.4069**	0.4770
EDA	0.4727	0.4053	0.4831
BERT Aug	0.4455	0.3692	0.4395
Backtranslation	0.4045	0.3451	0.4249

Tabel 9. Rangkuman Model Logictic Regression Skenario 3

Augmentation	*Accuracy*	*F1 Macro*	*F1 Weighted*
Modified EDA	0.4773	**0.3832**	0.4787
Backtranslation	0.4864	0.3817	0.4816
BERT Aug	0.4500	0.3745	0.4442
EDA	0.4727	0.3703	0.4733

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

Berdasarkan tabel-tabel hasil pada skenario 3, tabel berikut merangkum teknik augmentasi terbaik yang menghasilkan performa optimal untuk masing-masing model.

Tabel 10. Rangkuman Model Skenario 3 Terbaik

Model	***Best Augmentation***	***Accuracy***	***F1-Macro***	***Peningkatan***
BERT	***Original EDA***	**5.636**	**4.761**	**+0.0558**
DistilBERT	*Backtranslation*	5.545	4.462	+0.0617
Linear SVM	*Modified EDA*	4.682	3.809	+0.0051
Naive Bayes	*Modified EDA*	4.636	4.069	+0.2590
Logistic Regression	*Modified EDA*	4.773	3.832	+0.1436

Hasil eksperimen menunjukkan bahwa teknik augmentasi memberikan dampak positif yang signifikan, terutama pada metrik F1-Macro. BERT dengan *Original EDA* mencapai performa tertinggi secara keseluruhan (F1-Macro 0.4761). Selain itu, model-model klasik seperti *Linear Support Vector Machine*, *Naive Bayes*, dan *Logistic Regression* memperoleh kinerja terbaik ketika menggunakan *Modified EDA*, sedangkan model *Transformer* lebih responsif terhadap *Original EDA* dan *Backtranslation*. Secara keseluruhan, performa terbaik pada penelitian ini dihasilkan oleh *BERT* dengan augmentasi *EDA*. Adapun *DistilBERT* menunjukkan kinerja optimal saat dikombinasikan dengan *Backtranslation*, yang kemungkinan memberikan variasi kalimat yang lebih natural dan berbeda secara semantik dibandingkan sekadar substitusi kata sederhana.

3.2 Pembahasan

Berdasarkan hasil pengujian tiga skenario yang dilakukan sebelumnya, didapatkan hasil analisis sebagai berikut:

3.2.1 Analisis Skenario 1 (TF-IDF Baseline)

Baseline TF-IDF menunjukkan keterbatasan utama pada kemampuan pemodelan konteks. Representasi TF-IDF memetakan teks ke dalam bobot kata dan frasa berdasarkan frekuensi dan kekhasannya di dokumen. Pendekatan ini efektif untuk tugas topik atau klasifikasi yang kelasnya kontras, tetapi menjadi kurang optimal untuk klasifikasi rating 1–5 yang secara semantik berurutan dan tipis batasnya.

Linear SVM menjadi model terbaik dalam S1 karena sifatnya yang kuat pada ruang fitur berdimensi tinggi dan *sparse*, serta kemampuannya membangun *decision boundary* yang relatif stabil meski data terbatas. Logistic Regression menunjukkan performa menengah. Naive Bayes menjadi yang terendah kemungkinan karena asumsi independensi antar kata yang terlalu sederhana untuk menangkap pola sentimen kompleks pada ulasan restoran.

3.2.2 Analisis Skenario 2 (Transformer Baseline)

Peningkatan performa pada baseline Transformer menegaskan bahwa tugas klasifikasi rating membutuhkan pemahaman konteks dan urutan kata, bukan sekadar frekuensi kata. BERT memiliki keunggulan dalam menangkap dependensi antar kata melalui mekanisme *self-attention*, sehingga dapat mengenali pola evaluasi halus seperti perbandingan, kontras, atau intensitas sentimen.

DistilBERT, sebagai versi lebih ringkas dari BERT, menawarkan trade-off antara efisiensi dan performa. Hasil S2 memperlihatkan bahwa DistilBERT masih mampu mendekati kinerja BERT, sehingga layak dipertimbangkan ketika keterbatasan sumber daya komputasi menjadi faktor penting.

3.2.3 Analisis Skenario 3 (Augmentasi)

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

Augmentasi pada skenario ini memberikan dampak yang bervariasi antar model. Pada model klasik, Modified EDA cenderung menjadi strategi terbaik. Hal ini dapat dijelaskan karena augmentasi ringan berbasis substitusi/penyisipan kata masih mempertahankan distribusi kosakata yang kompatibel dengan representasi TF-IDF. Dengan kata lain, teks hasil augmentasi tetap "terlihat" mirip secara statistik bagi model berbasis *bag-of-words.*

Sebaliknya, augmentasi yang lebih berat seperti BERT augmentation dan backtranslation kadang menghasilkan kalimat dengan struktur lebih beragam. Walaupun secara semantik mungkin lebih kaya, variasi ini tidak selalu langsung menguntungkan model klasik karena perubahan distribusi kata dapat menggeser pola yang dipelajari TF-IDF.

Pada sisi Transformer, EDA dan backtranslation dapat memberikan manfaat lebih jelas. EDA membantu memperkaya variasi kata tanpa mengubah struktur kalimat terlalu ekstrem. Backtranslation berpotensi menghasilkan parafrase yang lebih natural, sehingga menambah keragaman konteks. Hal ini tampak pada performa DistilBERT yang meningkat paling baik pada backtranslation.

3.2.4 Mengapa 5 Label Cenderung Sulit Mencapai Akurasi Tinggi

Terdapat beberapa faktor utama yang menyebabkan model sulit mencapai akurasi mendekati 0.8 pada skenario klasifikasi rating 1–5:

1. **Batas antar kelas yang tipis dan bersifat ordinal**
   Rating 4 dan 5 sering kali memiliki ekspresi positif yang hampir serupa, begitu juga rating 2 dan 3 dalam nuansa netral–negatif. Selain itu, ulasan dengan nada positif tetap dapat berakhir pada rating 3 karena alasan lain seperti ekspektasi pengguna, harga, atau konteks pengalaman personal. Hal ini menyebabkan perbedaan antarkelas menjadi kurang jelas.
2. **Ukuran dataset yang relatif terbatas**
   Dengan jumlah data hanya sekitar seribuan ulasan, variasi gaya bahasa, panjang ulasan, serta konteks pengalaman pengguna menjadi sangat beragam. Untuk dapat mempelajari pola yang konsisten pada tiap kelas, model membutuhkan jumlah contoh yang lebih besar.
3. **Ketidakseimbangan distribusi kelas**
   Pada dataset ulasan restoran, rating tinggi umumnya lebih dominan. Kondisi ini mendorong model untuk lebih sering memprediksi kelas mayoritas, sehingga performa pada kelas minoritas menurun. Hal ini terlihat dari nilai F1-Macro yang lebih rendah dibandingkan F1-Weighted.
4. **Adanya *noise* dan subjektivitas dalam ulasan**
   Dua pengguna dapat menulis ulasan dengan isi yang hampir sama tetapi memberikan rating berbeda. Variasi subjektivitas seperti ini menghasilkan *label noise* yang secara alami membatasi performa maksimal model, terlepas dari algoritme yang digunakan.

3.2.5 Implikasi Hasil

Secara umum, hasil eksperimen menunjukkan beberapa temuan penting. Model Transformer konsisten memberikan kinerja yang lebih baik dibandingkan pendekatan berbasis TF-IDF pada tugas klasifikasi rating 1–5. Teknik augmentasi terbukti lebih efektif ketika diterapkan pada model Transformer, terutama augmentasi yang mempertahankan naturalitas dan koherensi kalimat. Sementara itu, untuk model klasik, augmentasi ringan seperti *Modified EDA* cenderung lebih aman dibandingkan teknik augmentasi yang menghasilkan perubahan besar pada struktur kalimat. Temuan ini mendukung asumsi bahwa klasifikasi rating bukan sekadar menilai keberadaan kata positif atau negatif, tetapi menuntut pemahaman terhadap intensitas, nuansa, dan konteks pengalaman pengguna yang tersampaikan dalam ulasan.

3.2.6 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang dapat memengaruhi performa model. Hal tersebut kami rincikan sebagai berikut yaitu:

1. Jumlah data yang relatif terbatas.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :

2. Potensi *label noise* akibat subjektivitas pengguna.
3. Penggunaan preprocessing yang cukup ketat, seperti pembersihan karakter dan lemmatization yang berpotensi menghilangkan sinyal ekspresif tertentu, misalnya penekanan, tanda seru, atau gaya bahasa informal.

3.2.7 Rekomendasi Pengembangan

Untuk meningkatkan performa pada klasifikasi 5 label, terdapat beberapa strategi yang dapat dipertimbangkan, antara lain:

1. Menerapkan pendekatan *ordinal classification* yang memodelkan rating sebagai variabel berurutan.
2. Menggunakan pendekatan regresi lalu melakukan pembulatan sebagai pembanding.
3. Menerapkan *class weighting* secara lebih konsisten pada seluruh model.
4. Menambahkan fitur tambahan jika tersedia, seperti metadata restoran atau kategori ulasan.
5. Memperluas jumlah data dengan sumber yang relevan untuk meningkatkan keragaman dan stabilitas model.

4. KESIMPULAN

Penelitian ini mengevaluasi performa model klasifikasi rating restoran berbasis ulasan teks dengan target lima kelas melalui tiga skenario pengujian. Berdasarkan hasil eksperimen, baseline Transformer menunjukkan performa yang lebih baik dibanding baseline TF-IDF. Pada skenario TF-IDF, Linear SVM menjadi model terbaik dengan F1-macro 0.3758, sedangkan Naive Bayes dan Logistic Regression menunjukkan performa yang lebih rendah. Hal ini mengindikasikan bahwa representasi berbasis frekuensi kata belum cukup untuk membedakan nuansa antar rating yang berdekatan.

Pada baseline Transformer, BERT dan DistilBERT menghasilkan peningkatan kinerja dibanding model klasik. BERT memberikan F1-macro yang lebih tinggi (0,4203) dibanding DistilBERT, menunjukkan bahwa representasi konteks yang lebih kaya masih memberi keuntungan pada dataset ini. Skenario augmentasi memberikan dampak yang lebih konsisten pada Transformer dibanding model klasik. Model terbaik dalam penelitian ini adalah BERT dengan augmentasi EDA, yang menghasilkan F1-macro tertinggi sebesar 0.4761 dengan akurasi 0.5636. Sementara itu, DistilBERT dengan backtranslation menjadi kombinasi terbaik kedua dengan F1-macro 0.4462. Temuan ini menegaskan bahwa augmentasi yang menghasilkan variasi kalimat tetap natural dan tidak terlalu mengganggu struktur makna sangat membantu dalam meningkatkan generalisasi model berbasis konteks.

Meskipun performa terbaik belum mencapai tingkat akurasi tinggi (misalnya ≥0,8), hasil penelitian ini sudah menunjukkan pola yang jelas bahwa klasifikasi rating lima kelas merupakan tugas yang inherently sulit akibat batas kelas yang tipis, potensi label noise, dan perbedaan subjektivitas pengguna. Penelitian selanjutnya dapat mengeksplorasi pendekatan ordinal classification, eksperimen regression-to-rating, penggunaan class weighting yang lebih seragam, serta penambahan data atau fitur kontekstual lain (misalnya metadata restoran) guna meningkatkan performa pada klasifikasi rating 1–5.

REFERENSI

[1] E. L. C. Macedo, F. C. Delicato, L. F. M. de Moraes, and G. Fortino, "Assigning Trust to Devices in the Context of Consumer IoT Applications," *IEEE Consumer Electronics Magazine*, vol. 13, no. 1, pp. 12–21, Jan. 2024, doi: 10.1109/MCE.2022.3154357.

[2] P. P. Putra, "Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 921–935, May 2025, doi: 10.47738/jads.v6i2.618.

[3] S. Feng *et al.*, "A Survey of Data Augmentation Approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Stroudsburg, PA, USA:

Association for Computational Linguistics, 2021, pp. 968–988. doi: 10.18653/v1/2021.findings-acl.84.

[4] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput Soc Netw*, vol. 8, no. 1, p. 1, Dec. 2021, doi: 10.1186/s40649-020-00080-x.

[5] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, "A comparative review on deep learning models for text classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, p. 325, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp325-335.

[6] F. Muftie and M. Haris, "IndoBERT Based Data Augmentation for Indonesian Text Classification," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, Aug. 2023, pp. 128–132. doi: 10.1109/ICITRI59340.2023.10250061.

[7] R. Shirke and A. Agrawal, "Performance analysis of token-based text augmentation techniques on text classification tasks in Indic languages," in *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, IEEE, Aug. 2023, pp. 168–172. doi: 10.1109/ICSCC59169.2023.10335009.

[8] . Natasya and A. S. Girsang, "Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis," *Emerging Science Journal*, vol. 7, no. 1, pp. 256–272, Nov. 2022, doi: 10.28991/ESJ-2023-07-01-018.

[9] S. Saranya and G. Usha, "A Machine Learning-Based Technique with Intelligent WordNet Lemmatize for Twitter Sentiment Analysis," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 339–352, 2023, doi: 10.32604/iasc.2023.031987.

[10] J. Wang, X. Li, J. He, Y. Zheng, and J. Ma, "Enhancing Implicit Sentiment Learning via the Incorporation of Part-of-Speech for Aspect-based Sentiment Analysis."

[11] R. Budiarto Hadiprakoso, H. Setiawan, R. Novita Yasa, and P. Siber dan Sandi Negara, "Text Preprocessing for Optimal Accuracy in Indonesian Sentiment Analysis Using a Deep Learning Model with Word Embedding."

[12] A. Maulana, Inayah Khasnaputri Afifah, Asghafi Mubarrak, Kiagus Rachmat Fauzan, Ardhan Dwintara, and B. P. Zen, "COMPARISON OF LOGISTIC REGRESSION, MULTINOMIALNB, SVM, AND K-NN METHODS ON SENTIMENT ANALYSIS OF GOJEK APP REVIEWS ON THE GOOGLE PLAY STORE," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1487–1494, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.863.

[13] N. Smairi, H. Abadlia, H. Brahim, and W. L. Chaari, "Fine-tune BERT based on Machine Learning Models For Sentiment Analysis," *Procedia Comput Sci*, vol. 246, pp. 2390–2399, 2024, doi: 10.1016/j.procs.2024.09.531.

[14] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, Dec. 2020, pp. 117–121. doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[15] L. Ruan and Q. Jin, "Survey: Transformer based video-language pre-training," *AI Open*, vol. 3, pp. 1–13, 2022, doi: 10.1016/j.aiopen.2022.01.001.

[16] G. G. Şahin, "To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP," *Computational Linguistics*, vol. 48, no. 1, pp. 5–42, Apr. 2022, doi: 10.1162/coli_a_00425.

[17] A. Devo, G. Costante, and P. Valigi, "Deep Reinforcement Learning for Instruction Following Visual Navigation in 3D Maze-Like Environments," *IEEE Robot Autom Lett*, vol. 5, no. 2, pp. 1175–1182, Apr. 2020, doi: 10.1109/LRA.2020.2965857.

[18] S. Kasus dan Implementasi Menggunakan Python Edisi, "MACHINE LEARNING Ibnu Daqiqil ID."