



PROPOSAL FINAL PROJECT

Peningkatan Kinerja *Fine-Grained Sentiment Analysis* pada Ulasan Restoran Menggunakan Pendekatan *Text Augmentation*

AUDREY SASQHIA WIJAYA

NRP 5025221055

AMANDA ILLONA FARREL

NRP 5025221056

Dosen Pengampu

Dini Adni Navastara, S.Kom., M.Sc.

NIP 198510172015042001

Program Studi Teknik Informatika

Departemen Teknik Informatika

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Surabaya

Tahun 2025

ABSTRAK

Peningkatan Kinerja *Fine-Grained Sentiment Analysis* pada Ulasan Restoran Menggunakan Pendekatan *Text Augmentation*

Abstrak

Analisis sentimen *fine-grained* pada ulasan restoran memberikan wawasan mendalam mengenai kepuasan pelanggan melalui skala rating yang spesifik (1 hingga 5). Namun, tantangan utama dalam pengembangan model klasifikasi yang akurat adalah ketersediaan dataset publik yang sering kali terbatas dan memiliki distribusi kelas yang sangat tidak seimbang (*imbalanced*). Kondisi ini menyebabkan model *Machine Learning* maupun *Deep Learning* cenderung bias terhadap kelas mayoritas dan mengalami *overfitting*, sehingga gagal melakukan generalisasi dengan baik pada kelas minoritas atau nuansa sentimen yang kompleks.

Sebagai solusi untuk mengatasi keterbatasan tersebut, penelitian ini menerapkan strategi *Text Data Augmentation* yang bertujuan memperkaya variasi data latih secara sintesis dengan tetap mempertahankan makna semantik asli. Metode augmentasi yang diterapkan meliputi *Original Easy Data Augmentation* (EDA), *Modified EDA* yang mengintegrasikan aturan *Part-of-Speech* (POS) *Tagging* untuk menjaga kata kunci sentimen, serta teknik *Backtranslation* (Inggris-Jerman) menggunakan model MarianTM. Data hasil augmentasi kemudian digunakan untuk melatih dan mengevaluasi kinerja lima model klasifikasi yang mewakili tingkat kompleksitas berbeda, yaitu *Naive Bayes*, *Linear SVM*, *Long Short-Term Memory* (LSTM), serta model berbasis *Transformer* (*BERT* dan *RoBERTa*).

Pengujian dilakukan melalui skenario eksperimental yang membandingkan performa model pada data *baseline* (tanpa augmentasi) terhadap data yang telah diberi perlakuan augmentasi. Evaluasi difokuskan pada pengukuran peningkatan metrik Akurasi dan *Macro F1-Score* untuk menilai kemampuan model dalam menangani ketidakseimbangan kelas. Penelitian ini bertujuan untuk mengidentifikasi teknik augmentasi yang paling efektif dalam meningkatkan kinerja generalisasi model pada tugas *fine-grained sentiment analysis*, serta menganalisis respons model *Transformer* modern terhadap intervensi data sintesis dibandingkan dengan model klasik.

Kata kunci:

Fine-Grained Sentiment Analysis, Text Data Augmentation, Modified EDA, Backtranslation, Transformer.

DAFTAR ISI

ABSTRAK	i
DAFTAR ISI	ii
DAFTAR GAMBAR	iii
DAFTAR TABEL	iv
DAFTAR SIMBOL	v
BAB 1 PENDAHULUAN	6
1.1 Latar Belakang	6
1.2 Rumusan Masalah	6
1.3 Batasan Masalah	7
1.4 Tujuan	7
1.5 Manfaat	7
BAB 2 TINJAUAN PUSTAKA	9
2.1 Hasil Penelitian Terdahulu	9
2.2 Dasar Teori	10
2.2.1 Fine-Grained Sentiment Analysis	10
2.2.2 Data Imbalance dan Low-Resource	10
2.2.3 Data Imbalance dan Low-Resource	11
2.2.4 Text Data Augmentation (TDA)	11
2.2.5 Algoritma Klasifikasi Teks	12
2.2.6 Metrik Evaluasi Klasifikasi	14
BAB 3 METODOLOGI	16
3.1 Lingkungan Kerja Sistem	16
3.2 Dataset yang Digunakan	16
3.3 Perancangan Sistem	17
3.4 Skenario Pengujian	18
DAFTAR PUSTAKA	19

DAFTAR GAMBAR

Gambar 2. 1 Rumus probabilitas yang digunakan Naïve Bayes (Elzeiny, 2023)	12
Gambar 2. 2 Ilustrasi SVM (Saxena, 2017)	12
Gambar 2. 3 Ilustrasi LSTM (CallMeTwitch, 2022)	13
Gambar 2. 4 Cara kerja BERT (Rath, 2023)	13
Gambar 2. 5 Cara kerja RoBERTa (Naveed, 2023)	13
Gambar 2. 6 Rumus Confusion Matrix (Irina, 2023)	14
Gambar 3. 1 Diagram Alir Proses (Dokumen Pribadi, 2025)	17

DAFTAR TABEL

Tabel 2. 1 State of Art

9

DAFTAR SIMBOL

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Ulasan pelanggan pada platform daring seperti Google Maps memegang peranan penting dalam pengembangan bisnis. Sebab, hal tersebut merupakan opini jujur mengenai berbagai aspek layanan seperti makanan, harga, dan suasana. Analisis sentimen pada ulasan dapat membantu pemilik bisnis memahami persepsi pelanggan. Namun, *fine-grained sentiment analysis* masih menghadapi tantangan berupa keterbatasan dan ketidakseimbangan data latih berlabel. Ketidakseimbangan ini menyebabkan model *Deep Learning* (DL) mengalami *overfitting* dan gagal melakukan generalisasi pada data baru.

Untuk mengatasi kelangkaan dan ketimpangan data, *Text Data Augmentation* (TDA) diterapkan sebagai strategi untuk meningkatkan keragaman dan kuantitas data latih tanpa perlu mengumpulkan data baru secara manual. Teknik augmentasi berbasis token yang populer adalah *Easy Data Augmentation* (EDA), yang meliputi penggantian sinonim, penyisipan acak, penghapusan acak, dan pertukaran kata (Muftie & Haris, 2023). Meskipun EDA terbukti efektif dan mudah diimplementasikan, teknik ini memiliki risiko mengubah makna kalimat atau menghilangkan konteks sentimen yang penting karena sifat operasinya yang acak. Sebagai solusi atas kelemahan augmentasi acak tersebut, penelitian terbaru mengajukan metode yang lebih terkontrol seperti *Modified EDA* yang memanfaatkan *Part-of-Speech* (POS) *Tagging* untuk menjaga kata-kata bermuatan sentimen agar tidak terhapus atau tertukar. Selain itu, teknik *Backtranslation* juga terbukti efektif dalam menghasilkan parafrasa yang alami dan mempertahankan makna asli ulasan (Natasya & Girsang, 2023).

Di sisi lain, pemilihan model klasifikasi juga penting. Model klasik seperti *Bi-Directional Long Short-Term Memory* (Bi-LSTM) dan *Convolutional Neural Network* (CNN) diketahui sangat diuntungkan oleh augmentasi data karena membutuhkan variasi data yang besar untuk belajar (Shirke & Agrawal, 2023). Sementara itu, model berbasis *Transformer* seperti BERT atau IndoBERT yang telah dilatih pada korpus besar memiliki pemahaman konteks yang kuat, sehingga dampaknya terhadap augmentasi data perlu diteliti lebih lanjut apakah memberikan peningkatan signifikan atau stagnan seperti yang ditemukan pada kasus IndicBERT.

Berdasarkan konteks tersebut, Penelitian ini bertujuan untuk mengeksplorasi efektivitas Augmentasi Data Teks pada dataset ulasan restoran berbahasa Inggris. Secara khusus, penelitian ini akan membandingkan dampak augmentasi terhadap dua kelompok model yang berbeda: model *Machine Learning* konvensional dan klasik (*Naive Bayes*, *SVM*, *LSTM*) yang biasanya membutuhkan banyak fitur manual, dibandingkan dengan model *Pre-trained Language Model* modern (*BERT*, *RoBERTa*) yang sudah memiliki pengetahuan kebahasaan yang luas.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana pengaruh penerapan teknik augmentasi *Original EDA*, *BERT Augmentation*, dan *Backtranslation* terhadap distribusi data dan keragaman linguistik pada kelas minoritas dalam dataset *fine-grained* ulasan restoran?

2. Bagaimana perbandingan respons kinerja antara model *Deep Learning* klasik (seperti CNN dan Bi-LSTM) dengan model berbasis *Transformer* (seperti BERT/RoBERTa) ketika dilatih menggunakan dataset yang telah diaugmentasi?
3. Apakah metode *BERT Augmentation* dan *Backtranslation* mampu menghasilkan peningkatan metrik evaluasi (Akurasi dan F1-Score) yang lebih signifikan dan stabil dibandingkan metode EDA standar dalam menangani klasifikasi ulasan berskala 1 sampai 5?

1.3 Batasan Masalah

Agar penelitian ini tetap terarah dan dapat diselesaikan dalam lingkup waktu yang ditentukan, batasan masalah yang ditetapkan adalah:

1. Objek Penelitian: Dataset yang digunakan adalah *Google Maps Restaurant Reviews* yang diambil dari repositori publik (Kaggle).
2. Lingkup Klasifikasi: Penelitian berfokus pada *Fine-Grained Sentiment Analysis* dengan target klasifikasi berupa rating skala 1 hingga 5.
3. Metode Augmentasi: Teknik augmentasi yang dibandingkan adalah *Original EDA*, *BERT Augmentation*, dan *Backtranslation*.
4. Model Klasifikasi: Model klasifikasi yang digunakan untuk evaluasi perbandingan kinerja meliputi model *Machine Learning* klasik (Naive Bayes, Linear SVM), model *Deep Learning* berbasis sekuens (LSTM), dan model berbasis *Transformer* (BERT, RoBERTa).
5. Metrik Evaluasi: Keberhasilan peningkatan kinerja diukur menggunakan metrik Akurasi (*Accuracy*) dan *Macro F1-Score* untuk menangani ketidakseimbangan kelas.

1.4 Tujuan

Untuk menjawab rumusan masalah yang telah diidentifikasi sebelumnya, penelitian ini bertujuan untuk:

1. Mengimplementasikan teknik EDA, *BERT Augmentation* dan *Backtranslation* sebagai upaya penanganan *imbalanced data* pada dataset ulasan restoran tanpa mengubah makna sentimen asli (*rating*).
2. Menganalisis karakteristik model *Deep Learning* klasik dan *Pre-trained Language Model* dalam merespons penambahan data sintetis.
3. Mengukur dan membuktikan tingkat peningkatan kinerja (persentase kenaikan Akurasi dan F1-Score) yang dihasilkan oleh masing-masing teknik augmentasi terhadap *baseline* (tanpa augmentasi) pada tugas *fine-grained sentiment analysis*.

1.5 Manfaat

Selain bertujuan untuk menjawab rumusan masalah yang telah ditetapkan, manfaat penelitian ini adalah sebagai berikut:

1. Menjadi referensi dalam pemilihan strategi penanganan data tidak seimbang yang optimal pada ulasan pelanggan dengan skala penilaian detail.
2. Memberikan perbandingan performa berbagai model, dari Naive Bayes hingga RoBERTa, terhadap teknik augmentasi data teks sebagai dasar pertimbangan efisiensi dan performa model.
3. Menjadi dasar pengembangan sistem analisis sentimen yang lebih akurat dalam membedakan tingkat kepuasan pelanggan sehingga mendukung pengambilan keputusan bisnis yang lebih tepat.

BAB 2 TINJAUAN PUSTAKA

2.1 Hasil Penelitian Terdahulu

Penelitian mengenai *Text Data Augmentation* (TDA) telah berkembang pesat sebagai solusi untuk mengatasi masalah keterbatasan data (*low-resource*) dan ketidakseimbangan kelas (*imbalanced class*) dalam *Natural Language Processing* (NLP).

Paper 1 meneliti augmentasi teks untuk Bahasa Indonesia menggunakan pendekatan berbasis model bahasa (*Language Model-based*). Mereka mengusulkan teknik *selective insertion* menggunakan IndoBERT. Hasil penelitian menunjukkan bahwa teknik berbasis IndoBERT mampu menghasilkan data sintetik yang mempertahankan makna dan sentimen asli, serta terbukti meningkatkan akurasi model CNN dan Bi-LSTM, mengungguli teknik augmentasi berbasis aturan sederhana seperti *Random Insert* (Muftie & Haris, 2023).

Sementara itu, Paper 2 melakukan analisis kinerja augmentasi berbasis token pada bahasa-bahasa India (*Indic languages*). Temuan krusial dari penelitian ini adalah adanya perbedaan dampak augmentasi terhadap jenis model yang digunakan. Augmentasi berbasis token (seperti EDA) memberikan peningkatan kinerja yang signifikan pada model klasik seperti Bi-LSTM. Namun, model *Transformer* yang telah dilatih secara masif (seperti IndicBERT) tidak mendapatkan manfaat signifikan dari augmentasi token sederhana karena model tersebut sudah memiliki pemahaman konteks yang kaya (Shirke & Agrawal, 2023).

Lebih spesifik pada domain ulasan restoran, Paper 3 mengusulkan metode hibrida yang menggabungkan *Modified EDA* dan *Backtranslation* untuk *Aspect-Based Sentiment Analysis* (ABSA). Penelitian ini menekankan bahwa teknik EDA standar (seperti *Random Swap/Delete*) berisiko merusak makna sentimen atau menghilangkan kata kunci aspek. Oleh karena itu, mereka memodifikasi EDA dengan aturan *Part-of-Speech* (POS) *tagging* dan menggabungkannya dengan *Backtranslation*. Metode hibrida ini terbukti meningkatkan *F1-Score* hingga 3,2% (Natasya & Girsang, 2023).

Dari rangkuman dan posisi penelitian ini terhadap penelitian terdahulu disajikan pada tabel berikut:

Tabel 2. 1 *State of Art*

Paper	Dataset	Metode	Model	Analisis Gap
IndoBERT Based Data Augmentation for Indonesian Text Classification (Muftie & Haris, 2023)	Twitter Text (Bahasa Indonesia)	IndoBERT <i>Selective Insertion</i> vs <i>Random Insert</i>	CNN, Bi LSTM	Temuan: Augmentasi berbasis PLM (IndoBERT) lebih unggul menjaga semantik dibanding <i>random insert</i> . Gap: Fokus hanya pada <i>insert</i> , belum membandingkan dengan <i>backtranslation</i> atau <i>modified swap/delete</i> .
Performance analysis of token-based text augmentation techniques on text classification tasks in Indic languages (Shirke & Agrawal, 2023)	Berita & Sentimen (Bahasa India)	Token-based EDA (<i>Synonym Replacement, Random Insertion, Random Swap, Random Deletion</i>)	Bi-LSTM, IndicBERT	Temuan: Model Transformer kurang mendapat manfaat dari augmentasi token sederhana dibanding Bi-LSTM. Gap: Belum mengeksplorasi augmentasi yang lebih canggih (<i>context-aware</i>) untuk model Transformer.

Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis (Natasya & Girsang, 2023)	Ulasan Restoran (Trip Advisor)	<i>Original EDA, Modified EDA, Backtranslation</i>	LSTM, Bi-LSTM, CNN	<p>Temuan: Modifikasi EDA dengan POS Tagging krusial untuk menjaga makna pada ulasan restoran.</p> <p>Gap: Fokus pada <i>Aspect-Based</i> (ABSA), belum diterapkan pada analisis <i>Fine-Grained</i> skala 1-5.</p>
--	--------------------------------	--	--------------------	---

Penelitian ini berfokus pada *dataset Google Maps Restaurant Reviews* berbahasa Inggris dengan target klasifikasi *fine-grained* (skala 1–5) yang memiliki tingkat kompleksitas lebih tinggi dibandingkan klasifikasi sentimen biner. Penelitian ini menerapkan pendekatan komparatif antara *Original EDA*, *Modified EDA*, dan *Backtranslation* (Inggris–Jerman) yang diuji pada berbagai model klasifikasi, yaitu Naive Bayes, SVM, LSTM, BERT, dan RoBERTa. Kebaruan penelitian ini terletak pada pengujian dampak strategi augmentasi data secara luas terhadap model klasik (Naive Bayes dan SVM) dibandingkan dengan model modern berbasis *Transformer* (BERT dan RoBERTa) pada data *fine-grained* yang tidak seimbang, guna melihat perbedaan respons kinerja antar kelompok model tersebut.

2.2 Dasar Teori

Subbab ini menyajikan landasan teori yang krusial bagi analisis dan pemodelan dalam penelitian ini.

2.2.1 Fine-Grained Sentiment Analysis

Analisis sentimen adalah proses komputasi untuk mengidentifikasi dan mengategorikan opini yang diungkapkan dalam teks. Berbeda dengan analisis sentimen biner yang hanya mengelompokkan teks ke dalam kelas Positif atau Negatif, *Fine-Grained Sentiment Analysis* memberikan tingkat granularitas yang lebih presisi, biasanya menggunakan skala Likert (misalnya, Sangat Negatif, Negatif, Netral, Positif, Sangat Positif) yang direpresentasikan dengan rating bintang 1 hingga 5. Tingkat kerumitan *fine-grained* lebih tinggi karena model harus mampu membedakan nuansa emosi yang tipis (misalnya, membedakan rating 4 “Bagus” dengan rating 5 “Sangat Bagus”).

2.2.2 Data Imbalance dan Low-Resource

Dalam konteks analisis data dan pembelajaran mesin, tahap persiapan data adalah fase fundamental yang memastikan kualitas dan kesiapan data sebelum proses analisis atau pemodelan yang lebih kompleks dilakukan. Proses ini secara umum melibatkan dua metode utama yang saling melengkapi, yakni Exploratory Data Analysis (EDA) dan Preprocessing Data. Kedua metode ini esensial untuk mengubah data mentah menjadi format yang optimal untuk analisis dan interpretasi yang akurat.

1. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan pendekatan investigatif dalam menganalisis kumpulan data untuk meringkas karakteristik utamanya, sering kali dengan bantuan visualisasi. Tujuan utama dari Exploratory Data Analysis (EDA) adalah untuk memahami struktur data, mengidentifikasi pola tersembunyi, mendeteksi *outlier*, menguji hipotesis awal, dan memeriksa asumsi dengan bantuan statistik deskriptif dan representasi grafis (Wickham & Grolemond, 2017).

Exploratory Data Analysis (EDA) juga mencakup sejumlah langkah teknis yang penting dalam proses pembersihan dan pemahaman data sebelum dilakukan analisis lebih lanjut. Langkah-langkah tersebut antara lain: menampilkan data, memahami proporsi dataset, mengidentifikasi nilai hilang, memeriksa tipe data, mendeteksi nilai duplikat, menangani nilai tidak konsisten, serta menganalisis distribusi data (Kalirane, 2024).

2. **Preprocessing Data**

Metode ini merupakan serangkaian teknik yang digunakan untuk mengubah data mentah ke dalam format yang bersih, konsisten, dan terstruktur agar siap digunakan dalam analisis atau pemodelan pembelajaran mesin. Data mentah sering kali mengandung masalah seperti ketidaklengkapan, inkonsistensi, atau *noise*, yang jika tidak ditangani dapat menurunkan kinerja model dan mengurangi keandalan hasil analisis (Géron, 2019).

Sejalan dengan itu, preprocessing data merupakan tahap krusial dalam *data pipeline* yang bertujuan meningkatkan kualitas data sebelum dianalisis lebih lanjut. Tahapan ini mencakup berbagai langkah, seperti mendeteksi dan menghapus *outliers*, mengonversi variabel kategorikal ke bentuk numerik melalui teknik *encoding*, serta menyeimbangkan distribusi data untuk menghindari bias dalam pemodelan. Pentingnya identifikasi nilai non-numerik, nilai tidak konsisten, dan ketidakseimbangan data juga ditekankan sebagai bagian dari proses pembersihan data yang komprehensif (Kalirane, 2024).

2.2.3 Data Imbalance dan Low-Resource

Masalah *low-resource* terjadi ketika jumlah data latih berlabel sangat terbatas, yang umum terjadi pada bahasa selain Bahasa Inggris, seperti Bahasa Indonesia. Masalah ini sering diperparah dengan kondisi *imbalanced class*, di mana distribusi data antar kelas tidak merata (misalnya, ulasan bintang 5 jauh lebih banyak daripada bintang 2). Hal ini menyebabkan model cenderung bias ke kelas mayoritas dan mengalami *overfitting*.

2.2.4 Text Data Augmentation (TDA)

Text Data Augmentation (TDA) adalah teknik meningkatkan jumlah dan variasi data latih secara sintesis tanpa mengumpulkan data baru secara manual.

1. **Easy Data Augmentation (EDA)**

Diperkenalkan oleh Wei dan Zou, EDA terdiri dari empat operasi acak: *Synonym Replacement* (SR), *Random Insertion* (RI), *Random Swap* (RS), dan *Random Deletion* (RD). Meskipun cepat, metode ini memiliki kelemahan dapat mengubah makna kalimat jika kata penting (seperti “tidak” atau “enak”) terhapus atau tertukar.

2. **Backtranslation**

Metode ini menerjemahkan teks dari bahasa asli (Indonesia) ke bahasa target (misal: Inggris atau Mandarin) dan kemudian menerjemahkannya kembali ke bahasa asli. Proses ini menghasilkan parafrasa yang memiliki struktur kalimat berbeda namun makna yang sama.

3. **BERT Augmentation**

Berbeda dengan EDA yang bergantung pada sinonim kamus, BERT Augmentation memanfaatkan model bahasa berbasis Transformer seperti BERT atau IndoBERT untuk menghasilkan variasi teks secara kontekstual. Secara umum, beberapa token dalam

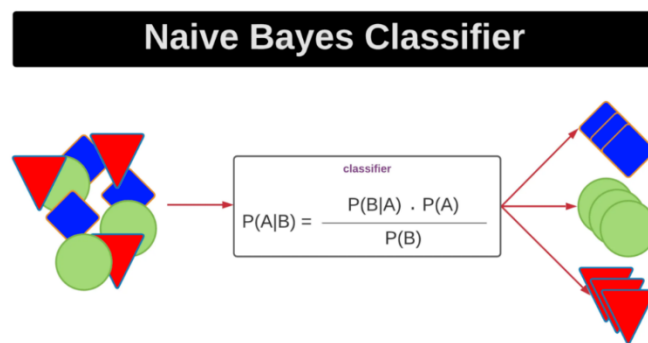
kalimat di-mask, lalu model memprediksi kata pengganti yang paling mungkin berdasarkan konteks kata-kata di sekelilingnya. Dengan demikian, kata yang diganti cenderung tetap selaras dengan makna kalimat secara keseluruhan, sehingga risiko perubahan polaritas sentimen (misalnya karena kata penting seperti “tidak”, “buruk”, atau “enak” terganti) lebih kecil dibanding EDA.

2.2.5 Algoritma Klasifikasi Teks

Penelitian ini menggunakan variasi algoritma untuk melihat dampak augmentasi pada berbagai tingkat kompleksitas model.

1. Naive Bayes

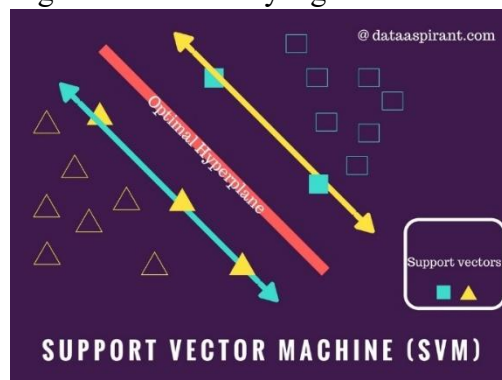
Metode klasifikasi probabilistik sederhana yang didasarkan pada Teorema Bayes dengan asumsi independensi yang kuat antar fitur. Meskipun sederhana, Naive Bayes sering digunakan sebagai *baseline* yang efektif untuk klasifikasi teks karena kecepatannya.



Gambar 2. 1 Rumus probabilitas yang digunakan Naïve Bayes (Elzeiny, 2023)

2. Support Vector Machine (SVM)

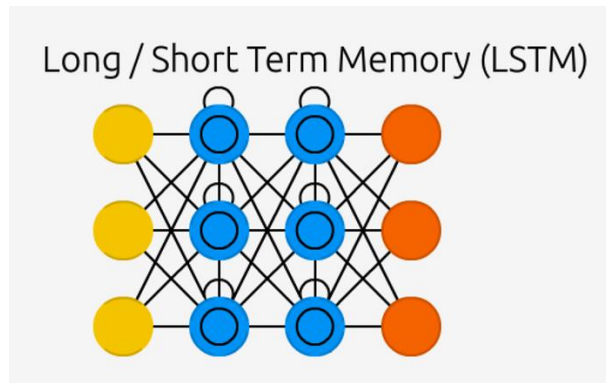
SVM adalah algoritma *supervised learning* yang bekerja dengan mencari *hyperplane* terbaik yang memisahkan kelas data dalam ruang vektor berdimensi tinggi. Pada klasifikasi teks, *Linear SVM* sering kali memberikan performa yang sangat baik, terutama pada dataset dengan dimensi fitur yang besar.



Gambar 2. 2 Ilustrasi SVM (Saxena, 2017)

3. Long Short-Term Memory (LSTM)

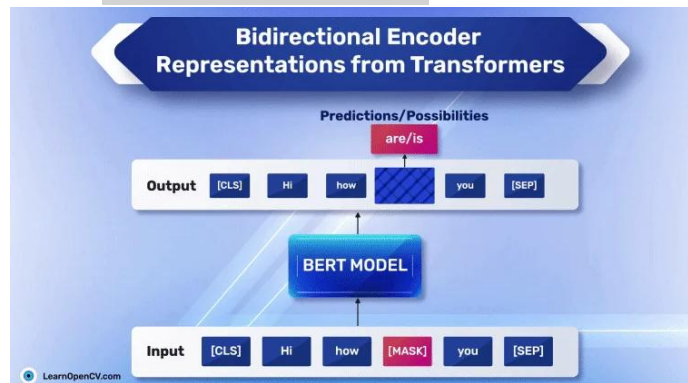
LSTM adalah jenis *Recurrent Neural Network* (RNN) yang mampu mempelajari ketergantungan jangka panjang (*long-term dependencies*) dalam urutan teks, mengatasi masalah *vanishing gradient* yang sering terjadi pada RNN standar.



Gambar 2. 3 Ilustrasi LSTM (CallMeTwitch, 2022)

4. BERT (*Bidirectional Encoder Representations from Transformers*)

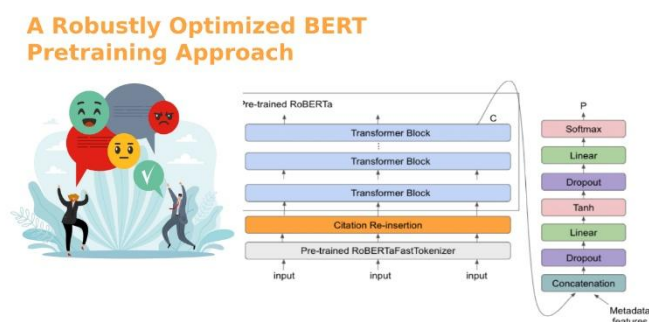
BERT adalah model berbasis *Transformer* yang dilatih menggunakan teknik *Masked Language Modeling* (MLM). Berbeda dengan model yang membaca teks secara urut (kiri ke kanan), BERT membaca seluruh urutan kata sekaligus (*bidirectional*), memungkinkannya memahami konteks kata yang ambigu dengan sangat baik. Model yang digunakan adalah `bert-base-uncased`.



Gambar 2. 4 Cara kerja BERT (Rath, 2023)

5. RoBERTa (*Robustly Optimized BERT Pretraining Approach*)

RoBERTa adalah varian optimasi dari BERT. Perbedaan utamanya terletak pada proses pelatihan: RoBERTa menghapus tugas *Next Sentence Prediction* (NSP), menggunakan data latih yang jauh lebih besar, dan melatih model dengan durasi yang lebih lama, sehingga sering kali menghasilkan performa yang lebih tinggi dibanding BERT standar.



Gambar 2. 5 Cara kerja RoBERTa (Naveed, 2023)

2.2.6 Metrik Evaluasi Klasifikasi

Metrik evaluasi adalah tolok ukur kuantitatif yang esensial untuk menilai performa, akurasi, dan tingkat kesalahan dari sebuah model. Karena penelitian ini merupakan tugas klasifikasi (*classification task*), maka evaluasi dilakukan menggunakan metrik berikut

1. *Confusion Matrix*

Tabel representasi yang membandingkan prediksi model dengan label sebenarnya, terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Gambar 2. 6 Rumus *Confusion Matrix* (Irina, 2023)

Kelebihannya adalah memberikan gambaran detail mengenai jenis kesalahan yang dibuat oleh model. Kita dapat melihat secara spesifik kelas mana yang sering "tertipu" dengan kelas lain (misalnya, membedakan apakah model sering salah memprediksi Rating 2 sebagai Rating 1). Adapun kekurangannya Adalah tidak memberikan satu angka skor tunggal untuk perbandingan antar model secara cepat. Pada kasus klasifikasi banyak kelas (*multi-class*) seperti rating 1-5, tabel menjadi besar (5x5) dan lebih sulit dianalisis secara sekilas dibandingkan skor sederhana.

2. Akurasi (*Accuracy*)

Akurasi merupakan rasio jumlah prediksi yang benar terhadap total seluruh prediksi (Scikit-learn developers, 2025). Berikut rumusnya:

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

Keterangan variabel:

- *TP (True Positive)* merupakan jumlah data positif yang diprediksi dengan benar sebagai positif oleh model.
- *TN (True Negative)* merupakan jumlah data negatif yang diprediksi dengan benar sebagai negatif oleh model.
- *FP (False Positive)* merupakan jumlah data negatif yang salah diprediksi sebagai positif (Kesalahan Tipe I).
- *FN (False Negative)* merupakan jumlah data positif yang salah diprediksi sebagai negatif (Kesalahan Tipe II).

Kelebihannya adalah sangat intuitif dan mudah dipahami oleh orang awam. Adapun kekurangannya Adalah sangat sensitif terhadap data yang tidak seimbang. Jika data didominasi oleh kelas mayoritas (misal: Rating 5), model yang hanya memprediksi Rating 5 terus-menerus tetap akan memiliki akurasi tinggi, padahal model tersebut gagal mengenali kelas minoritas (Rating 1 atau 2).

3. *F1-Score (Macro-Average)*

F1-Score adalah rata-rata harmonis dari *Precision* dan *Recall*. Untuk data tidak seimbang, varian **Macro-F1** sangat penting karena menghitung F1 untuk setiap kelas (rating 1-5) secara terpisah lalu merata-ratakannya, memberikan bobot yang sama untuk kelas minoritas dan mayoritas (Scikit-learn developers, 2025).

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

Dimana:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Keterangan variabel:

- *Precision* merupakan tingkat ketepatan prediksi positif model (seberapa banyak dari prediksi positif yang benar-benar positif).
- *Recall* merupakan tingkat keberhasilan model menemukan kembali data positif (seberapa banyak data positif asli yang berhasil dideteksi).
- *TP, FN, FP* Sama seperti definisi pada metrik Akurasi.

Ini merupakan metrik terbaik untuk dataset yang tidak seimbang (*imbalanced*). *Macro-Average* memastikan bahwa performa pada kelas minoritas (yang datanya sedikit) tetap dianggap sama pentingnya dengan kelas mayoritas. Namun, interpretasinya tidak se-intuitif akurasi. Nilai F1-Score membutuhkan pemahaman mengenai konsep *precision* dan *recall* untuk dapat dimaknai dengan benar.

BAB 3 METODOLOGI

3.1 Lingkungan Kerja Sistem

Proses pengolahan data dan pelatihan model dilakukan menggunakan beberapa library utama. Library pandas dan numpy digunakan untuk memanipulasi dan mengelola data tabular, sedangkan matplotlib dimanfaatkan untuk visualisasi data. Scikit-learn membantu dalam proses pembagian data menjadi data latih, validasi, dan uji serta untuk melakukan perhitungan metrik evaluasi seperti akurasi, presisi, recall, F1 score, dan confusion matrix.

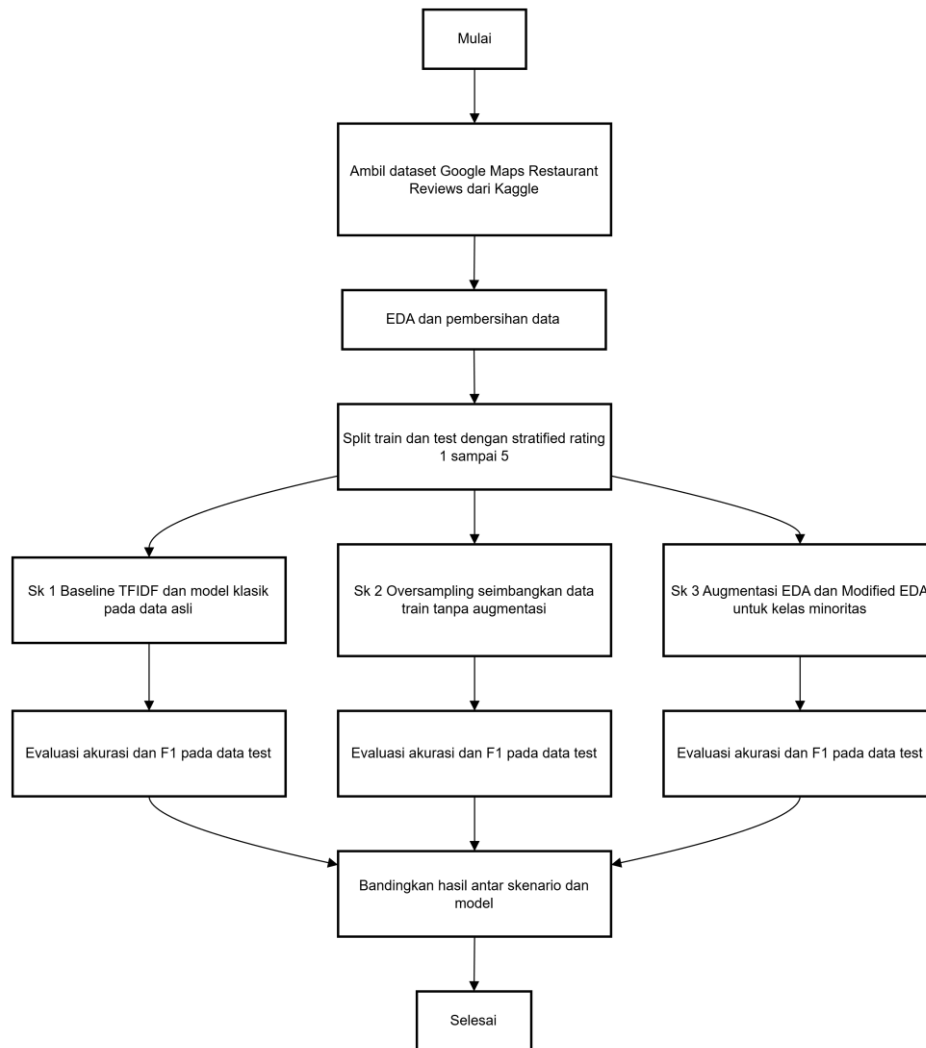
Untuk pemodelan berbasis deep learning, penelitian ini menggunakan PyTorch yang nantinya dikombinasikan dengan transformers untuk memanggil model pre-trained seperti roberta-base atau bert-base-uncased. Pada sisi augmentasi teks, NLTK dengan modul WordNet digunakan untuk memperoleh sinonim dalam implementasi teknik Easy Data Augmentation, sedangkan spaCy digunakan untuk melakukan penandaan kelas kata pada Modified EDA. Selain itu, untuk melakukan proses backtranslation dari bahasa Inggris ke bahasa Jerman menggunakan model MarianTM.

3.2 Dataset yang Digunakan

Dataset yang digunakan dalam penelitian ini adalah Google Maps Restaurant Reviews yang didapatkan dari Kaggle (Bilgin, 2023). Dataset ini berisi sekitar 1100 ulasan restoran dalam bahasa Inggris, dengan fokus pada review pelanggan di berbagai restoran. Atribut dalam dataset meliputi:

- restaurant_name: nama restoran yang diulas
- review: ulasan dalam bahasa Inggris
- rating: rating bintang dalam skala 1 sampai 5 yang diberikan oleh pengguna
- Terdapat beberapa atribut lain seperti date, link, serta beberapa metadata tambahan yang tidak digunakan untuk tugas ini.

3.3 Perancangan Sistem



Gambar 3. 1 Diagram Alir Proses (Dokumen Pribadi, 2025)

Tahap awal dimulai dengan pengambilan dataset Google Maps Restaurant Reviews dari Kaggle yang berisi teks ulasan dan rating bintang 1–5. Data mentah ini kemudian melalui proses pembersihan data (data cleaning) untuk menghapus baris yang memiliki nilai kosong, memastikan dataset yang digunakan konsisten dan bebas error sebelum masuk ke tahap selanjutnya.

Setelah data bersih, dilakukan pembagian dataset menjadi data latih, validasi, dan uji dengan strategi stratified split untuk menjaga proporsi distribusi rating. Data uji diperlakukan secara khusus tanpa diberi perlakuan augmentasi agar hasil evaluasi kelak merefleksikan kemampuan generalisasi model pada data nyata. Tahap berikutnya adalah preprocessing teks minimalis, mengingat model Transformer memiliki tokenizer yang tangguh; proses ini hanya meliputi normalisasi sederhana seperti penghapusan spasi berlebih agar format teks siap diproses.

Pada tahap augmentasi, alur kerja bercabang menjadi beberapa varian skenario untuk memperkaya data latih. Selain menggunakan data asli sebagai baseline, diterapkan teknik Original EDA (penggantian sinonim, penyisipan, penukaran, dan penghapusan kata), Modified

EDA (pengembangan EDA dengan aturan POS tagging untuk menjaga aspek sentimen), serta Backtranslation (penerjemahan bolak-balik untuk parafrasa natural). Teknik-teknik ini bertujuan menambah variasi dan jumlah data pada kelas minoritas.

Setiap varian dataset yang terbentuk kemudian digunakan untuk melatih model klasifikasi berbasis Transformer. Teks ulasan diproses melalui tokenizer internal model sebelum digunakan dalam pelatihan untuk memprediksi rating. Selama proses ini, data validasi digunakan sebagai acuan untuk memantau kinerja dan memilih model terbaik dari setiap skenario (berdasarkan metrik seperti akurasi atau F1-score), memastikan model yang dipilih adalah yang paling optimal sebelum masuk ke tahap pengujian akhir.

Tahap terakhir adalah evaluasi komparatif menggunakan data uji murni yang tidak pernah disentuh proses augmentasi. Performa model dari skenario tanpa augmentasi, Original EDA, Modified EDA, dan Backtranslation dibandingkan untuk melihat dampak masing-masing teknik. Dari perbandingan ini, ditarik kesimpulan mengenai efektivitas setiap metode augmentasi dalam meningkatkan kualitas prediksi model Transformer terhadap sentimen ulasan restoran.

3.4 Skenario Pengujian

Pada skenario pertama, setelah dilakukan eksplorasi data dan preprocessing dasar, label target dipertahankan sesuai rating asli (skala 1 hingga 5) untuk mendukung analisis *fine-grained*. Model kemudian dilatih terlebih dahulu pada data latih dengan distribusi asli (tanpa penambahan data sintetis) menggunakan kelima model klasifikasi yang digunakan dalam penelitian. Kinerja yang dihasilkan pada skenario ini berfungsi sebagai baseline dan sekaligus menunjukkan dampak nyata dari ketidakseimbangan kelas (*class imbalance*) pada setiap tingkat rating, sehingga menjadi dasar argumen bahwa penambahan data untuk kelas minoritas sangat diperlukan

Pada skenario kedua jika hasil eksplorasi menunjukkan bahwa jumlah ulasan positif dan negatif tidak sepenuhnya seimbang maka akan dilakukan penambahan data tapi tanpa melibatkan augmentasi. Pada skenario ini, tiga model (LSTM, Linear SVM, dan Naive Bayes) serta dua model pretrained (BERT dan RoBERTa) dilatih menggunakan data latih asli tanpa augmentasi dan dievaluasi pada himpunan data uji yang sama. Hasil yang diperoleh digunakan untuk melihat karakteristik dan kekuatan relatif masing-masing model pada dataset ulasan restoran, serta menetapkan model-model mana yang layak dijadikan acuan pada skenario berikutnya.

Skenario ketiga menitikberatkan pada analisis pengaruh augmentasi terhadap kinerja model. Lalu setelah data di augmentasi melakukan perbandingan antar modelnya dengan data yang sudah di augmentasi untuk melihat apakah model lebih bagus setelah dilakukan augmentasi. Model augmentasi yang digunakan adalah EDA, Modified EDA, dan backtranslation. Setiap model yang dihasilkan kembali dievaluasi pada data uji yang sama. Perbandingan akurasi dan F1-score antar varian ini memungkinkan untuk menilai apakah augmentasi benar-benar memberikan peningkatan kinerja, serta apakah Modified EDA yang lebih terkontrol dapat menghasilkan performa yang lebih stabil dibanding EDA klasik, baik pada model klasik maupun pada model pretrained.

DAFTAR PUSTAKA

- Bilgin, D. (2023). *Google Maps Restaurant Reviews*. Retrieved from Kaggle: <https://www.kaggle.com/datasets/denizbilginn/google-maps-restaurant-reviews/data>
- CallMeTwitch. (2022, December 19). *Building a Neural Network Zoo from scratch: The Long Short-Term Memory Network*. Retrieved from Medium: <https://medium.com/@CallMeTwitch/building-a-neural-network-zoo-from-scratch-the-long-short-term-memory-network-1cec5cf31b7>
- Elzeiny, M. (2023, June 18). *The Ultimate Guide to Naive Bayes*. Retrieved from Machine Learning Archive: <https://mlarchive.com/machine-learning/the-ultimate-guide-to-naive-bayes/>
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol: O'Reilly Media, Inc.
- Irina, E. (2023, June 11). *Memahami Confusion Matrix, Accuracy, Precision, Recall, Specificity, dan F1-Score*. Retrieved from Medium: <https://esairina.medium.com/memahami-confusion-matrix-accuracy-precision-recall-specificity-dan-f1-score-610d4f0db7cf>
- Kalirane, M. (2024, April 4). *EDA vs Data Preprocessing: What's the Difference?* Retrieved June 24, 2025, from Analytics Writer: <https://analyticswriter.com/eda-data-preprocessing/>
- Muftie, F., & Haris, M. (2023). IndoBERT Based Data Augmentation for Indonesian Text Classification. *2023 International Conference on Information Technology Research and Innovation (ICITRI)* (pp. 1-5). Jakarta: IEEE. doi:10.1109/ICITRI59340.2023.10250061
- Natasya, & Girsang, A. S. (2023, February). Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis. *Emerging Science Journal*, 7, 256-272. doi:10.28991/ESJ-2023-07-01-018
- Naveed, F. (2023, October 27). *A Robustly Optimized BERT Pretraining Approach - Hands-on using Python*. Retrieved from Farrukh Naveed Blog: <https://farrukhnaveed.co/blogs/a-robustly-optimized-bert-pretraining-approach-hands-on-using-python/>
- Rath, S. R. (2023, October 31). *BERT - Bidirectional Encoder Representations from Transformers*. Retrieved from LearnOpenCV: <https://learnopencv.com/bert-bidirectional-encoder-representations-from-transformers/>
- Saxena, R. (2017, January 13). *Svm classifier, Introduction to support vector machine algorithm*. Retrieved from Data Aspirant: <https://dataaspirant.com/support-vector-machine-algorithm/>
- Scikit-learn developers. (2025). *sklearn.metrics.accuracy_score*. Retrieved from Scikit-learn documentation: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score

- Scikit-learn developers. (2025). *sklearn.metrics.f1_score*. Retrieved from Scikit-learn documentation: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
- Shirke, R., & Agrawal, A. (2023). Performance analysis of token-based text augmentation techniques on text classification tasks in Indic languages. *International Conference on Smart Computing and Communications (ICSCC)* (pp. 1-5). India: IEEE. doi:10.1109/ICSCC59169.2023.10335009
- Wickham, H., & Grolemund, G. (2017). *R for Data Science*. Sebastopol: O'Reilly Media, Inc.