PERFORMANCE COMPARISON OF BERT METRICS AND CLASSICAL MACHINE LEARNING MODELS (SVM, NAIVE BAYES) FOR SENTIMENT ANALYSIS

PERBANDINGAN KINERJA METRIK BERT DAN MODEL MACHINE LEARNING KLASIK (SVM, NAIVE BAYES) UNTUK ANALISIS SENTIMEN

Adib Ulinuha El Majid[1], Reflan Nuari[2]
[1,2] Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: adib_ulinuha_el_majid@teknokrat.ac.id[1], reflan@teknokrat.ac.id[2]

Abstract *- Sentiment analysis is one of the important methods in understanding public opinion from large amounts of text, such as product reviews or user comments. Many studies have shown that the BERT (BiDirectional Encoder Representations from Transformers) model has advantages over classical machine learning models such as Support Vector Machine (SVM) and Naïve Bayes. However, there are still few studies that systematically compare the performance of the two on datasets from various topics and languages, especially those with imbalanced label distributions. This study compares four BERT variants (bert-base-uncased, distilbert-base-uncased, indobert-base-uncased, and distilbert-base-indonesian) with two classical models using three datasets of IMDb 50K (English), Amazon Food Reviews (English), and Gojek App Review (Indonesian). The classical model uses the TF-IDF vectorisation method, while the BERT model is optimised through a further training process (fine-tuning) with a layer freezing technique. The evaluation is carried out using accuracy, precision, recall, and F1-score. The results show that the BERT model excels on English data, while on imbalanced Indonesian data, SVM and Naïve Bayes produce higher F1-score results. These findings indicate that the selection of the right model must be adjusted to the characteristics of the data used.*
Keywords *- Sentiment Analysis, BERT, SVM, Naïve Bayes, IndoBERT, Performance Metrics*

Abstrak - Analisis sentimen menjadi salah satu metode penting dalam memahami opini publik dari teks dalam jumlah besar, seperti ulasan produk atau komentar pengguna. Banyak penelitian menunjukkan bahwa model BERT (*Bidirectional Encoder Representations from Transformers*) memiliki keunggulan dibandingkan model pembelajaran mesin klasik seperti *Support Vector Machine* (SVM) dan *Naïve Bayes*. Namun, masih sedikit studi yang secara sistematis membandingkan performa keduanya pada kumpulan data dari berbagai topik dan bahasa, terutama yang memiliki distribusi label tidak seimbang. Penelitian ini membandingkan empat varian BERT (bert-base-uncased, distilbert-base-uncased, indobert-base-uncased, dan distilbert-base-indonesian) dengan dua model klasik menggunakan tiga dataset IMDb 50K (bahasa Inggris), Amazon *Food Reviews* (bahasa Inggris), dan Gojek *App Reviews* (bahasa Indonesia). Model klasik menggunakan metode vektorisasi TF-IDF, sedangkan model BERT dioptimalkan melalui proses pelatihan lanjutan (fine-tuning) dengan teknik *layer freezing*. Evaluasi dilakukan menggunakan akurasi, presisi, recall, dan F1-score. Hasil menunjukkan bahwa model BERT unggul pada data berbahasa Inggris, sementara pada data berbahasa Indonesia yang tidak seimbang, SVM dan Naïve Bayes memberikan hasil F1-score lebih tinggi. Temuan ini menunjukkan bahwa pemilihan model yang tepat harus disesuaikan dengan karakteristik data yang digunakan.

Kata Kunci - Analisis Sentimen, BERT, SVM, Naïve Bayes, IndoBERT, Metrik Kinerja

I. PENDAHULUAN

Analisis sentimen telah menjadi metode krusial dalam mengolah ledakan data tekstual di era digital, memungkinkan pemahaman opini publik dari sumber seperti ulasan produk dan media sosial[1]. Sebagai salah satu cabang penting dalam Pemrosesan Bahasa Alami (*Natural Language Processing*/NLP), proses ini bertujuan mengidentifikasi, mengekstrak, dan mengelompokkan opini atau emosi dari suatu teks untuk mengetahui pandangan terhadap suatu isu atau objek[2]. Pendekatan awal dalam analisis sentimen seringkali mengandalkan model machine learning klasik seperti *Support Vector Machine* (SVM) dan *Naïve Bayes*. Model-model ini menawarkan dasar yang kuat dan telah terbukti efektif dalam banyak kasus, namun terkadang memiliki keterbatasan dalam menangkap nuansa kontekstual bahasa yang kompleks.

Kemajuan signifikan dalam pemrosesan bahasa alami (NLP) hadir dengan diperkenalkannya model berbasis *Transformer*, khususnya BERT (*Bidirectional Encoder Representations from Transformers*)[3]. BERT merevolusi pemahaman bahasa melalui representasi kontekstual dua arah dan kemampuan adaptasi (*fine-tuning*) yang efektif untuk berbagai tugas NLP, termasuk analisis sentimen. Berbagai studi, seperti yang dilakukan oleh Nugroho dkk. [4], telah menunjukkan potensi varian BERT spesifik bahasa seperti IndoBERT untuk mengungguli model klasik pada data lokal berbahasa Indonesia.

Meskipun demikian, tinjauan literatur menunjukkan adanya beberapa celah penelitian yang perlu diatasi. Banyak studi yang membandingkan BERT dengan model klasik cenderung berfokus pada satu jenis dataset atau satu bahasa, sehingga generalisasi kinerjanya pada skenario yang beragam masih terbatas. Selain itu, masih kurang perbandingan sistematis yang melibatkan berbagai varian BERT (termasuk versi ringan seperti DistilBERT dan versi spesifik bahasa seperti IndoBERT) melawan model klasik yang telah dioptimalkan (misalnya, dengan vektorisasi TF-IDF) pada kombinasi dataset multi-bahasa (Inggris dan Indonesia), multi-domain (ulasan film, makanan, aplikasi), dan dengan karakteristik keseimbangan kelas yang berbeda. Evaluasi kinerja pada kondisi data tidak seimbang, yang umum ditemui pada ulasan dunia nyata terutama untuk data berbahasa Indonesia, juga seringkali belum menjadi fokus utama, pemilihan metrik evaluasi yang komprehensif dan mampu mencerminkan performa sebenarnya pada data tersebut menjadi sangat krusial.

Penelitian ini bertujuan melakukan perbandingan kinerja secara sistematis antara empat varian BERT (bert-base-uncased, distilbert-base-uncased, indobert-base-uncased, distilbert-base-indonesian) dengan model klasik SVM dan Naïve Bayes. Evaluasi dilakukan pada tiga dataset representatif: IMDb Movie Reviews (Inggris, seimbang), Amazon Food Reviews (Inggris, tidak seimbang), dan Gojek App Reviews (Indonesia, tidak seimbang). Secara spesifik, penelitian ini akan mengeksplorasi bagaimana perbandingan kinerja (akurasi, presisi, recall, F1-score) antara model BERT dan model klasik pada dataset berbahasa Inggris dengan karakteristik keseimbangan kelas yang berbeda. Studi ini juga akan mendalami perbandingan kinerja, khususnya dalam metrik F1-score, antara model BERT dan model klasik pada dataset berbahasa Indonesia yang memiliki distribusi label tidak seimbang.

II. SIGNIFIKANSI STUDI

Studi ini menanggapi celah dalam literatur yang masih terbatas dalam membandingkan model *transformer-based* seperti BERT dengan model pembelajaran mesin klasik dalam konteks analisis sentimen. Beberapa penelitian sebelumnya jarang menyajikan perbandingan langsung antar pendekatan berbasis pembelajaran klasik dan model berbasis *deep learning* modern.

Hasil penelitian ini dapat memberikan acuan bagi praktisi *data science*, pengembang aplikasi, dan pelaku industri teknologi informasi dalam memilih model analisis sentimen yang paling sesuai dengan kebutuhan dan sumber daya mereka. Jika BERT terbukti secara signifikan lebih akurat, maka hal ini dapat mendorong adopsi teknologi *transformer* untuk aplikasi-aplikasi seperti monitoring opini publik, analisis *review* produk, dan pengelolaan media sosial. Sebaliknya, jika model klasik masih kompetitif, maka organisasi dengan keterbatasan komputasi dapat tetap menggunakan metode tersebut secara efisien.

A. Penelitian Terdahulu

Penelitian sebelumnya yang melakukan komparasi antara model BERT dan model kercerdasan buatan lainnya yang ditunjukan pada tabel berikut.

Tabel I
Penelitian Terdahulu

No	Penulis	Penelitian Terdahulu
1	Lenggo Geni[5]	Analisis Sentimen Tweet Sebelum Pemilu 2024 di Indonesia Menggunakan Model Bahasa IndoBERT. Menggunakan 4 varian model IndoBERT dengan akurasi mencapai sebesar 83,5% unggul hingga 14,49% dibandingkan model machine learning.[5]
2	Kuncahyo Setyo Nugroho[4]	BERT Fine-Tuning untuk Analisis Sentimen pada Ulasan Aplikasi Seluler Indonesia. Pengujian model varian model BERT-Base multilingual dan IndoBERT-Base dengan akurasi mecapai 82,29% unggul 3,39% dibandingan dibandingkan model machine learning.[4]
3	Muammar Khadapi[6]	Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024. Dimana model LSTM mencapai akurasi sebesar 85% sedangkan model BERT hanya mencapai akurasi sebesar 64,53%.[6]

Pada Tabel I merangkum beberapa penelitian terdahulu yang melakukan komparasi antara model BERT dan pendekatan lainnya. Misalnya, studi oleh Geni [5] menunjukkan keunggulan IndoBERT untuk analisis sentimen pemilu di Indonesia, sementara Nugroho dkk. [4] juga menyoroti kinerja baik IndoBERT pada ulasan aplikasi seluler. Di sisi lain, penelitian Khadapi [6] menemukan hasil yang berbeda dimana LSTM mengungguli BERT pada dataset spesifik mereka.

B. Tinjauan Pustaka

1. Analisis Sentimen

Analisis sentimen (*sentiment analysis*) atau *opinion mining* merupakan proses untuk memahami, mengekstraksi, dan mengolah data tekstual secara otomatis guna memperoleh informasi mengenai sentimen yang terkandung dalam suatu kalimat opini. Proses ini dilakukan untuk mengetahui pandangan atau kecenderungan opini seseorang terhadap suatu isu atau objek[2] Tujuan utama dari analisis sentimen adalah untuk mengidentifikasi apakah

suatu pernyataan dalam teks bersifat positif, negatif, atau netral. Kategori netral biasanya menunjukkan bahwa tidak ada opini yang jelas diungkapkan. Proses analisis ini dapat dilakukan pada berbagai tingkatan, termasuk tingkat dokumen secara keseluruhan, tingkat kalimat, maupun tingkat fitur atau aspek tertentu dalam teks[7].

2. BERT

Pada penelitian ini mengunakan model BERT, DistilBERT, dan varian *pre-train* BERT yaitu IndoBERT. Pemilihan model ini dimaksudkan dengan tujuan mebandingkan kualitas model dalam beradaptasi dengan dataset yang akan dilatih.

a. BERT

Model transformer ini telah melalui pelatihan awal menggunakan korpus besar teks berbahasa Inggris melalui metode *self-supervised learning*. Artinya, model dilatih hanya dengan teks mentah tanpa pelabelan manual oleh manusia. Sebagai gantinya, proses otomatis digunakan untuk menghasilkan input dan label dari data teks tersebut, memungkinkan pemanfaatan data publik dalam jumlah besar. Model dilatih terlebih dahulu pada BookCorpus, kumpulan data yang terdiri dari 11.038 buku yang belum diterbitkan dan Wikipedia bahasa[3].

b. DistilBERT

Model transformer yang lebih kecil dan lebih cepat daripada BERT, yang dilatih sebelumnya pada korpus yang sama dengan cara *self-supervised learning*, menggunakan model dasar BERT sebagai pengajar. Model dilatih sebelumnya pada data yang sama dengan BERT, yaitu BookCorpus, kumpulan data yang terdiri dari 11.038 buku yang belum diterbitkan dan Wikipedia bahasa Inggris[8].

c. IndoBERT

IndoBERT merupakan versi Bahasa Indonesia dari model BERT yang dilatih menggunakan lebih dari 220 juta kata yang dikumpulkan dari tiga sumber utama, yaitu Wikipedia Bahasa Indonesia (sekitar 74 juta kata), artikel berita dari Kompas, Tempo (Tala et al., 2003), dan Liputan6 (total sekitar 55 juta kata), serta korpus web berbahasa Indonesia (Indonesian Web Corpus) yang dikembangkan oleh Medved dan Suchomel (2017) dengan jumlah sekitar 90 juta kata[9].

3. SVM

Support Vector Machine (SVM) merupakan salah satu metode *supervised learning* yang digunakan dalam analisis data untuk keperluan pengenalan pola dan klasifikasi. Metode ini pertama kali diperkenalkan oleh Vapnik pada tahun 1992 dan bekerja berdasarkan prinsip *Structural Risk Minimization* yang bertujuan untuk meminimalkan kesalahan generalisasi. SVM bekerja dengan mencari hyperplane optimal yang dapat memisahkan data ke dalam kelas-kelas yang berbeda dengan margin terbesar[10].

4. Naïve Bayes

Naive Bayes merupakan salah satu metode klasifikasi yang sederhana dan didasarkan pada penerapan Teorema Bayes, dengan asumsi adanya independensi antar fitur. Metode ini menghitung probabilitas suatu kelas berdasarkan frekuensi data yang tersedia dalam basis data. Dalam konteks klasifikasi, tujuan utamanya adalah memprediksi label kelas dari suatu sampel berdasarkan sejumlah fitur atau karakteristik. Sebagai contoh, berdasarkan informasi

seperti usia, jenis kelamin, dan pendapatan, sistem dapat digunakan untuk memprediksi apakah seseorang cenderung memiliki pekerjaan dengan gaji tinggi atau tidak[11].

5. *Accuracy, Precision, Recall, F1-Score*

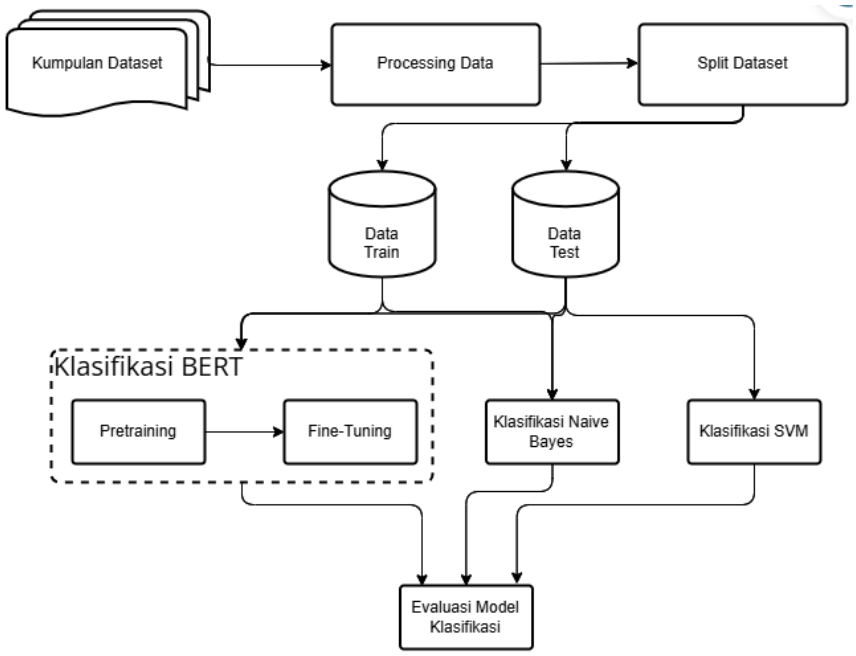
Penilaian kinerja model klasifikasi dalam studi ini menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-*score* guna menilai efektivitas model dalam melakukan klasifikasi sentimen.

Accuracy = (TP + TN) / (TP + FP + FN + TN)
Precision = (TP) / (TP + FP)
Recall = (TP) / (TP + FN)
F1 *Score* = 2 * (Recall * Precision) / (Recall + Precision)

Keterangan:
TP (*True Positive*): Jumlah data positif yang berhasil diprediksi dengan benar
TN *(True Negative*): Jumlah data negatif yang berhasil diprediksi dengan benar
FP (*False Positive*): Jumlah data negatif yang salah diprediksi sebagai positif
FN (*False Negative*): Jumlah data positif yang salah diprediksi sebagai negatif

C. Metode Penelitian

Penelitian ini terdiri dari beberapa tahapan. Pada tahap awal, setiap dataset melalui proses *Processing* Data, kemudian dilakukan pemisahan dataset menjadi data *training*, *testing*, dan *validation*. Data tersebut digunakan sebagai bahan *training* untuk setiap model yang akan dievaluasi. Berikut ini adalah tahapan-tahapan penelitian yang dilakukan.



Gambar 1. Metode Penelitian

1. Kumpulan Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari situs Kaggle, yaitu IMDB 50K, Amazon *Food Reviews*, serta satu dataset berbahasa Indonesia, yaitu *Gojek App Reviews* (Bahasa Indonesia). Ketiga dataset tersebut diimpor menggunakan bahasa pemrograman Python dan selanjutnya diproses lebih lanjut pada tahap data processing.

2. *Processing* Data
   Setelah berhasil diimpor, dataset akan melalui tahap data *processing*, yaitu proses pengolahan data agar dapat digunakan oleh model dalam proses pelatihan (*training*). Tahapan ini mencakup pembersihan data, normalisasi, serta konversi data teks ke dalam format yang dapat dibaca oleh model.

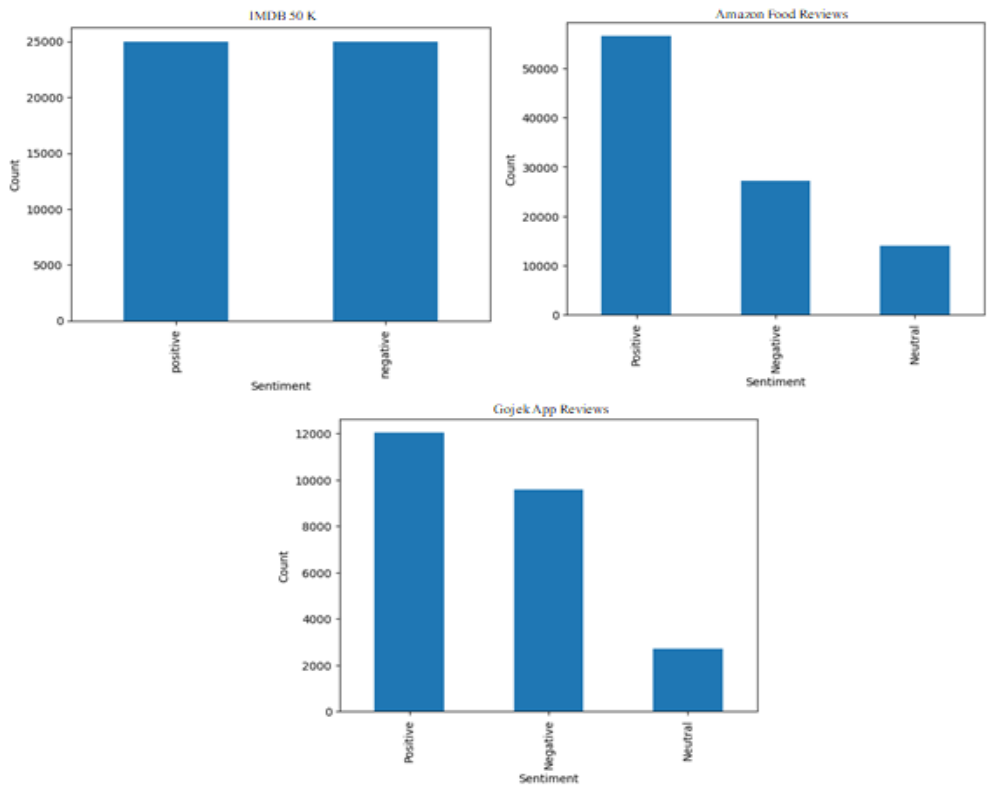
3. *Split Dataset*
   Setelah dataset diolah di *processing* data, dataset akan dipisah menjadi dua bagian yaitu data *train* yang akan digunakan model untuk melakukan training atau perubahan parameter pada model, dan data *test* yang berguna untuk menilai peforma model setelah melakukan *training*.

4. Klasifikasi Model
   Pada tahap ini, model klasifikasi diterapkan untuk melakukan analisis sentimen terhadap *dataset*. Model yang digunakan adalah model SVM, Naïve Bayes dan model berbasis *transformer*, yaitu BERT dan variannya. Model ini dilatih menggunakan data *training*, disesuaikan berdasarkan data *validation*, dan kemudian diuji performanya pada data *test*. Proses ini mencakup pelatihan model, penyetelan hiperparameter, dan optimalisasi hasil klasifikasi.

5. Evaluasi Model
   Setelah pelatihan selesai, model dievaluasi untuk mengukur efektivitasnya dalam melakukan klasifikasi sentimen. Evaluasi dilakukan menggunakan metrik-metrik umum seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Keempat metrik ini memberikan gambaran menyeluruh mengenai kinerja model, baik dari segi ketepatan klasifikasi maupun kemampuan dalam mendeteksi kelas yang relevan secara konsisten.

III. HASIL DAN PEMBAHASAN

A. Kumpulan Dataset
   Dataset yang digunakan dalam penelitian ini memiliki perbedaan dalam hal ukuran, jenis label, dan distribusi label. Dataset IMDB 50K terdiri dari dua label sentimen, yaitu positif dan negatif, dengan distribusi yang seimbang (rasio 1:1). Sementara itu, dataset Amazon *Food Review* dan Gojek *App Reviews* Bahasa Indonesia menggunakan sistem penilaian berupa skor dari 1 hingga 5. Untuk keperluan analisis sentimen, skor tersebut dikonversi menjadi tiga kategori label, yakni: skor 1 hingga 2 diklasifikasikan sebagai *negatif*, skor 3 sebagai *netral*, dan skor 4 hingga 5 sebagai *positif*. Setelah proses kategorisasi tersebut, kedua dataset menunjukkan distribusi label yang tidak seimbang.



Gambar 2. Distribusi Dataset IMDB 50k, Amazon Food Review, Gojek App Review

Dataset IMDb terdiri dari 50.000 data dengan distribusi label yang seimbang, yaitu masing-masing 25.000 data untuk sentimen positif dan negatif. Dataset Amazon Food Review, dengan total 97.717 data, dengan distribusi label yang tidak seimbang, terdiri dari 56.500 data positif, 27.221 data negatif, dan 13.996 data netral. Sementara itu, dataset Gojek App Review berjumlah 24.320 data, juga dengan distribusi tidak seimbang, yang mencakup 12.032 data positif, 9.588 data negatif, dan 2.700 data netral.

B. *Processing* Data

Proses ini melalui beberapa tahapan dengan tujuan akhir untuk mempersiapkan data agar dapat digunakan dalam pelatihan model secara optimal. Teknik data mining pada tahap ini mengubah data mentah menjadi format yang dapat dimengerti dengan cara membersihkan komentar tanpa opini, serta menghilangkan data yang tidak valid, mengandung noise, atau bersifat redundan[12] Tahapan data processing yang dilakukan dalam penelitian ini meliputi:

1. Pembersihan Data (*Data Cleaning*):

Pada tahap ini, dilakukan penghapusan karakter-karakter yang tidak relevan seperti simbol, angka, HTML *tags*, dan tanda baca yang tidak diperlukan. Selain itu, dilakukan normalisasi teks, termasuk pengubahan seluruh huruf menjadi huruf kecil (*lowercasing*) agar konsisten dalam analisis.

2. Penghapusan *Stopwords*:

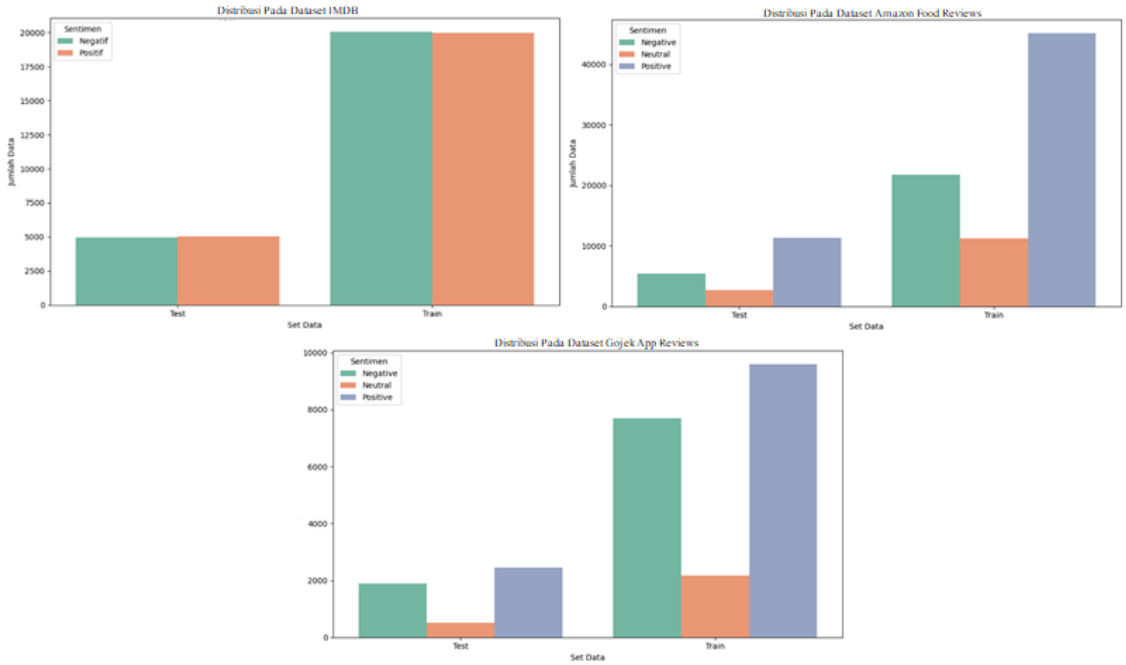
Stopwords adalah kata-kata umum yang tidak memiliki makna signifikan dalam analisis sentimen, seperti “yang”, “dan”, atau “adalah”. Kata-kata ini dihapus menggunakan daftar *stopwords* yang sesuai dengan bahasa pada masing-masing dataset (bahasa Inggris dan bahasa Indonesia).

3. *Stemming*/*Lemmatization*:

Proses *stemming* dan *lemmatization* dilakukan untuk menyederhanakan kata menjadi bentuk dasarnya guna meningkatkan konsistensi data. Stemming menghapus imbuhan pada kata, sedangkan lemmatization mengacu pada bentuk dasar sesuai konteks. Tahapan ini membantu mengurangi variasi kata yang tidak relevan dalam analisis sentimen.

C. Split Dataset

Data dibagi menjadi dua bagian, yaitu 80% untuk pelatihan dan 20% untuk pengujian, dengan proporsi yang tetap agar hasil pembagian dapat direproduksi secara konsisten. Pembagian ini dilakukan terhadap data ulasan sebagai input dan label sentimen sebagai target. Setelah proses ini, jumlah data pada masing-masing bagian diperiksa untuk memastikan distribusi yang sesuai.



Gambar 3. Pembagian Dataset Train dan Test Dataset IMDB 50K

D. Klasifikasi Model

Tahap klasifikasi merupakan inti dari analisis sentimen dalam penelitian ini, di mana beberapa model diterapkan untuk memprediksi sentimen berdasarkan data yang telah diproses sebelumnya. Tujuan dari tahap ini adalah untuk merancang arsitektur model serta menyetel *hyperparameter* agar diperoleh performa klasifikasi yang optimal. Model yang digunakan terdiri dari algoritma berbasis pembelajaran mesin tradisional seperti *Support Vector Machine* (SVM) dan Naïve Bayes, serta model berbasis *deep learning* modern, yaitu BERT dan variannya. Masing-masing model dilatih menggunakan data pelatihan dandiuji performanya pada data pengujian guna mengevaluasi efektivitas pendekatan yang digunakan.

1. Naïve Bayes dan SVM

Support Vector Machine (SVM) adalah metode klasifikasi dalam supervised learning yang dikenal karena dasar matematisnya yang kuat dan jelas. SVM bekerja dengan menemukan *hyperplane* optimal, yaitu fungsi pemisah antar kelas, melalui maksimasi jarak di antara kelas-kelas tersebut[13] Pada kedua model ini, dilakukan proses tokenisasi terhadap dataset untuk mengubah teks menjadi vektor numerik yang dapat diproses oleh model klasifikasi. Teknik tokenisasi yang digunakan adalah *TF-IDF Vectorization* dengan jumlah fitur maksimum sebanyak 5.000. Dalam bidang pengambilan informasi dan penambangan teks, metode TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan untuk tujuan

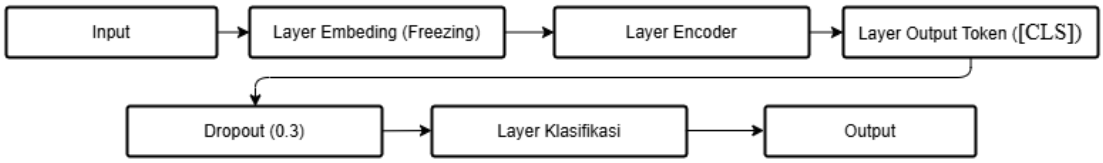
vektorisasi atau ekstraksi fitur teks, mengubahnya menjadi format vektor.[14] Setelah proses tokenisasi, dilakukan penyetelan hyperparameter menggunakan metode *GridSearchCV* guna memperoleh kombinasi parameter terbaik yang menghasilkan performa model paling optimal.

Tabel II
Hasil *Hyperparameter Tunning*

No	Model	Hyperparameter	Parameter Terbaik		
			IMDB	Amazon	Gojek
1	Naïve Bayes	'alpha': [0.1, 0.5, 1.0, 2.0, 5.0, 10]	0.5	0.1	2.0
2	SVM	'C': [0.01, 0.1, 1, 10]	0.1	0.1	0.1

2. BERT

Dalam penelitian ini digunakan model BERT dan variannya, yaitu DistilBERT, serta model pra-latih BERT yang dirancang khusus untuk Bahasa Indonesia, yaitu IndoBERT dan IndoDistilBERT. Sebelum data dimasukkan ke dalam model, dilakukan proses tokenisasi menggunakan tokenizer khusus BERT yang secara otomatis menambahkan *special token* seperti [CLS] di awal kalimat dan token penutup di akhir kalimat. Untuk keperluan klasifikasi, arsitektur BERT dimodifikasi dengan menambahkan *fully connected layer* (lapisan klasifikasi) di atas output token [CLS], yang berfungsi untuk menghasilkan prediksi kelas sentimen. Selain itu, dilakukan teknik *layer freezing* pada bagian *embedding* agar parameter-parameter tersebut tidak diperbarui selama pelatihan. Setelah struktur model disiapkan, dilakukan proses *fine-tuning*, yaitu pelatihan lanjutan pada model pra-latih menggunakan data yang spesifik dari penelitian ini agar model dapat menyesuaikan diri dengan karakteristik dataset dan meningkatkan akurasi dalam tugas klasifikasi sentimen. Pendekatan ini bertujuan untuk mengurangi waktu pelatihan, menghemat sumber daya komputasi, dan mencegah *overfitting*, terutama saat jumlah data pelatihan terbatas.



Gambar 4. Arsitektur Model Klasifikasi BERT

Tabel III
Parameter Model BERT

No	Parameter	bert-base-uncased	distilbert-base-uncased	indobert-base-uncased	distilbert-base-indonesian
1	Optimizer	AdamW	AdamW	AdamW	AdamW
2	Learning Rate	3e-5	3e-5	3e-5	3e-5
3	Batch Size	32	32	32	32
4	Max Length	128	128	128	128
5	Epoch	5	5	5	5

E. Evaluasi Model

Evaluasi dilakukan untuk menilai efektivitas masing-masing model dalam mengklasifikasikan sentimen dari teks ulasan pada berbagai dataset. Tujuan evaluasi adalah untuk memvalidasi kinerja sistem analisis sentimen dengan cara membandingkan hasil yang diberikannya terhadap kenyataan atau data aktual.[15] Proses ini melibatkan pengukuran performa model dengan menggunakan metrik umum seperti akurasi, presisi, recall, dan F1-score, guna memperoleh gambaran menyeluruh terkait kemampuan klasifikasi. Setiap model diuji pada data *test* yang belum pernah dilihat sebelumnya, sehingga dapat diketahui sejauh mana model mampu melakukan generalisasi terhadap data baru. Evaluasi ini juga mencerminkan bagaimana karakteristik masing-masing model, baik tradisional seperti SVM dan Naïve Bayes maupun model berbasis deep learning seperti BERT dan variannya, beradaptasi terhadap keragaman bahasa dan distribusi label.

Tabel IV
Evaluasi Model Pada Dataset IMDB 50K

IMDB 50K					
No	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
1	Naïve Bayes	0.8534	0.8573	0.8505	0.8539
2	SVM	0.8891	0.8784	**0.9051**	0.8916
3	bert-base-uncased	**0.8932**	**0.9025**	0.8835	**0.8929**
4	distilbert-base-uncased	0.8869	0.8882	0.8873	0.8877

Tabel V
Evaluasi Model Pada Dataset Amazon Food Reviews

Amazon Food Reviews					
No	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
1	Naïve Bayes	0.7180	0.6953	0.7180	0.6589
2	SVM	0.7659	0.7416	**0.7659**	0.7326
3	bert-base-uncased	**0.8276**	**0.7560**	0.7415	**0.7482**
4	distilbert-base-uncased	0.8146	0.7385	0.7478	0.7424

Tabel VI
Evaluasi Model Pada Dataset Gojek App Reviews Bahasa Indonesia

Gojek App Reviews Bahasa Indonesia					
No	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
1	Naïve Bayes	**0.7826**	0.7095	**0.7826**	**0.7406**
2	SVM	0.78	**0.7599**	0.78	0.7387
3	indobert-base-uncased	0.7819	0.6877	0.5858	0.5505
4	distilbert-base-indonesian	0.7862	0.6534	0.5913	0.5609

Pada dataset Gojek App Reviews yang tidak seimbang dan berbahasa Indonesia, model machine learning klasik SVM dan Naïve Bayes menunjukkan F1-score yang secara signifikan lebih unggul dibandingkan varian IndoBERT. Keunggulan ini dapat diatribusikan pada beberapa faktor kombinatif. Efektivitas vektorisasi TF-IDF yang digunakan oleh model klasik dalam menangkap dan memberikan bobot tinggi pada kata-kata kunci yang mungkin sangat indikatif untuk kelas minoritas dalam ulasan aplikasi, di mana pendekatan berbasis frekuensi kata mungkin kurang

terpengaruh dibandingkan model Transformer yang mencoba memahami konteks mendalam namun bisa terganggu oleh bahasa non-standar.

Perlu ditekankan bahwa hasil kinerja yang disajikan dalam Tabel IV, V, dan VI berasal dari satu kali proses evaluasi. Oleh karena itu, perbandingan antar model bersifat deskriptif dan observasional. Tanpa pengulangan eksperimen atau penggunaan teknik seperti *k-fold cross-validation* untuk menghasilkan beberapa sampel kinerja, tidak dimungkinkan untuk melakukan pengujian statistik formal (seperti *t-test* atau ANOVA) untuk menentukan apakah perbedaan yang diamati signifikan secara statistik.

IV. KESIMPULAN

Penelitian ini secara sistematis membandingkan kinerja empat varian BERT dengan model klasik SVM dan *Naïve Bayes* dalam tugas analisis sentimen, menggunakan tiga dataset dengan karakteristik bahasa, ukuran, dan keseimbangan kelas yang berbeda, sehingga secara dapat menjawab rumusan masalah terkait perbandingan kinerja pada skenario yang beragam. Hasil empiris menunjukkan bahwa model-model berbasis BERT umumnya menunjukkan keunggulan signifikan pada dataset berbahasa Inggris (IMDB dan Amazon *Food Reviews*), baik pada kondisi label seimbang maupun tidak seimbang, dalam berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-*score*. Sebaliknya, pada dataset Gojek *App Reviews* (Bahasa Indonesia), yang memiliki jumlah data yang lebih terbatas (kurang dari 25 ribu ulasan, berbeda signifikan dengan dataset IMDB dan Amazon *Food Review*) dan menunjukkan distribusi label yang sangat tidak seimbang setelah proses kategorisasi skor ulasan menjadi tiga label (positif, negatif, netral), model klasik SVM dan *Naïve Bayes* yang menggunakan vektorisasi TF-IDF justru menghasilkan F1-score yang lebih tinggi dibandingkan varian IndoBERT. Temuan krusial ini secara langsung menjawab tujuan penelitian untuk mendalami perbandingan kinerja, khususnya dalam metrik F1-score, pada dataset berbahasa Indonesia yang memiliki tantangan spesifik berupa ukuran data yang lebih kecil dan ketidakseimbangan kelas yang ekstrem ini, dan mengindikasikan bahwa model klasik lebih efektif dalam menyeimbangkan presisi dan recall dalam kondisi tersebut.

REFERENSI

[1] S. Mustafa, M. Fadhli, and D. R. Amanda, "Sistem Informasi Pemeliharaan Aset Elektronik Menggunakan SMS Gateway pada Dinas Pariwisata dan Kebudayaan Aceh," *Jurnal Nasional Komputasi dan Teknologi Informasi*, vol. 2, no. 2, 2019.

[2] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2012.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[4] K. S. Nugroho, A. Y. Sukmadewa, H. W. DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," Jul. 2021, doi: 10.1145/3479645.3479679.

[5] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 3, pp. 746–757, Aug. 2023, doi: 10.26555/jiteki.v9i3.26490.

[6] M. Khadapi *et al.*, "Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024," *Jurnal Komputer dan Informatika*, vol. 6, no. 2, pp. 130–137, Nov. 2024.

[7] W. Lei, K. Khine, N. Thwet, T. Aung, and T. T. Zin, "Feature Extraction Method for Aspect-Based Sentiment Analysis."

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[9] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.00677

[10] A. Kowalczyk, *Support vector machines succinctly*. 2017.

[11] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, pp. 1–7, Jan. 2023, doi: 10.35134/komtekinfo.v10i1.330.

[12] A. N. Syafia, M. F. Hidayattullah, and W. Suteddy, "Studi Komparasi Algoritma SVM Dan Random Forest Pada Analisis Sentimen Komentar Youtube BTS," vol. 8, no. 3, 2023.

[13] R. Pratiwi, D. D. S. H, and D. Af'idah, "Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM)," *INISTA*, vol. 4, pp. 40–6, Nov. 2021, doi: 10.20895.

[14] R. Kumar Bania, "COVID-19 Public Tweets Sentiment Analysis using TF-IDF and Inductive Learning Models." [Online]. Available: https://covid19.who.int/

[15] O. I. Gifari, M. Adha, I. Rifky Hendrawan, F. Freddy, and S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY*, vol. 2, no. 1, 2022.