Bayesian Tuning terhadap Model Pre-Trained PEGASUS untuk Peringkas Teks Informatif Berbahasa Indonesia

Kadek Artha Darma Pradnyana[*1], I Nyoman Prayana Trisna[2], Wayan Oger Vihikan[3]
Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Jl. Raya Kampus Unud Jimbaran, Badung 80361, Bali, Indonesia
Email: [1]artadarma37@gmail.com, [2] prayana.trisna@unud.ac.id, [3]oger_vihikan@unud.ac.id

Abstract. *This research explores abstractive text summarization of Indonesian news by fine-tuning the PEGASUS model using Bayesian optimization and enriched contextual inputs. The dataset contains 286,277 document-summary pairs scraped from JPNN.com, including titles and keyphrases used to construct informative input. Each section is marked with special tokens such as <TITLE>, <KEYPHRASES>, and <ARTICLE>. Evaluation using ROUGE and BERTScore shows that informative input substantially improves performance: +16.75% (ROUGE-1), +27.25% (ROUGE-2), +18.58% (ROUGE-L and ROUGE-Lsum), and +2.7% (BERTScore-F1) compared with regular input. Saliency analysis also shows consistently high sentence weights for contextual input components. Additionally, Bayesian hyperparameter tuning via Optuna yields marginal gains (+1.21% ROUGE-1, +2.1% ROUGE-2, +1.38% ROUGE-L & ROUGE-Lsum, +0.23% BERTScore) due to a limited number of trials (12) and a constrained hyperparameter search space. These findings demonstrate the effectiveness of contextual input design and the potential of Bayesian tuning to improve Transformer-based summarization for low-resource languages.*
Keywords: *Abstractive Text Summarizer, PEGASUS, Bayesian Optimization, Formatted Input, Informative Input.*

Abstrak. *Penelitian ini mengeksplorasi peringkasan teks abstraktif untuk berita berbahasa Indonesia dengan melakukan fine-tuning pada model PEGASUS menggunakan Bayesian Optimization dan input kontekstual yang diperkaya. Dataset berisi 286.277 pasangan dokumen–ringkasan yang diambil dari JPNN.com, lengkap dengan judul dan kata kunci yang digunakan untuk membentuk input informatif. Evaluasi menggunakan ROUGE dan BERTScore menunjukkan peningkatan substansial dari informative input: +16.75% (ROUGE-1), +27.25% (ROUGE-2), +18.58% (ROUGE-L & ROUGE-LSUM), dan +2.7% (BERTScore-F1) dibandingkan dengan input reguler. Analisis saliency menunjukkan bobot kalimat kontekstual yang konsisten tinggi. Penerapan hyperparameter tuning Bayesian melalui Optuna memberikan kenaikan marginal (+1.21% ROUGE-1, +2.1% ROUGE-2, +1.38% ROUGE-L & ROUGE-LSUM, +0.23% BERTScore) yang dipengaruhi oleh jumlah trial terbatas (12) dan ruang pencarian yang sempit. Temuan ini menegaskan efektivitas desain input kontekstual dan potensi hyperparameter tuning untuk peringkasan berbasis Transformer pada bahasa dengan sumber daya terbatas.*
Kata Kunci: *Peringkas Teks Abstraktif, PEGASUS, Optimasi Bayesian, Input Terformat, Input Informatif.*

1. Pendahuluan

Di era digital, volume informasi berkembang pesat dan diperkirakan berlipat ganda setiap dua tahun, melebihi kapasitas manusia untuk memprosesnya [1]. Di sisi lain, penelitian juga menemukan penurunan rentang perhatian rata-rata manusia dalam beberapa dekade terakhir [2]. Kombinasi antara kelebihan informasi dan penurunan kapasitas perhatian meningkatkan kebutuhan akan alat yang mampu menyajikan ringkasan singkat dan bernilai tinggi agar pembaca dapat menangkap gagasan utama secara efisien.

Berita tetap menjadi medium utama publik dalam mengonsumsi informasi, dan di Indonesia sebagian besar pembaca masih memilih format teks dalam mengonsumsi berita [3]. Portal nasional seperti JPNN.com menerbitkan artikel panjang dengan beragam topik (politik,

ekonomi, hukum, hiburan, olahraga), sehingga korpus berita menjadi target yang representatif sekaligus menantang untuk penelitian peringkasan otomatis yang diarahkan pada pembaca lokal.

Peringkasan otomatis berupaya mengompresi dokumen menjadi ringkasan singkat dan koheren sambil mempertahankan gagasan utama [4]. Dua paradigma utama eksis: pendekatan ekstraktif memilih dan menggabungkan fragmen teks yang penting [5], sedangkan pendekatan abstraktif menghasilkan kalimat baru yang mensintesis isi sumber, menghasilkan ringkasan yang lebih natural namun memerlukan kapabilitas generatif yang lebih tinggi [6], [7].

Arsitektur Transformer merevolusi penelitian peringkasan sejak diperkenalkan [8]. Model seperti BART, T5, dan khususnya PEGASUS mencapai kinerja unggul melalui mekanisme *self-attention* dan objektif pra-pelatihan yang selaras dengan tugas peringkasan [9]. Objektif GSG (Gap Sentence Generation) pada PEGASUS secara eksplisit dirancang untuk tugas peringkasan dan menunjukkan hasil yang kuat pada metrik ROUGE dan BERTScore [10]. Kinerja juga cenderung meningkat jika pra-pelatihan dan *fine-tuning* dilakukan dalam bahasa yang sama [11]. Varian PEGASUS berbahasa Indonesia tersedia untuk *fine-tuning* (thonyyy/pegasus_indonesian_base-finetune).

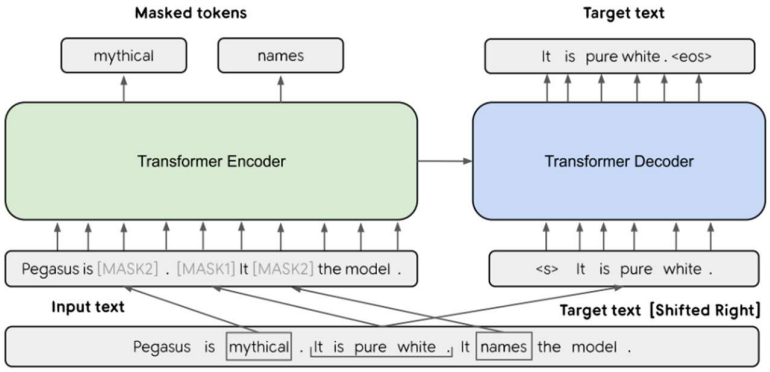
Penelitian ini menguji dua strategi komplementer untuk meningkatkan performa PEGASUS berbahasa Indonesia tanpa mengubah arsitektur model. Strategi pertama adalah memperkaya input dengan konteks terstruktur berupa judul dan kata kunci, mengikuti pendekatan input informatif yang terbukti membantu peringkasan berita Indonesia [12]. Strategi kedua adalah menerapkan Bayesian *hyperparameter tuning* untuk mengoptimalkan pilihan pelatihan seperti *learning rate*, *dropout*, dan *weight decay* [13], [14]. Eksperimen dinilai menggunakan metrik otomatis standar seperti ROUGE dan BERTScore untuk mengkuantifikasi peningkatan performa dari penambahan informasi kontekstual dan optimisasi *hyperparameter*.

2. Tinjauan Pustaka

2.1. PEGASUS

PEGASUS (*Pre-training with Extracted Gap-sentences for Abstractive Summarization*) adalah model pembelajaran mendalam *sequence-to-sequence* yang dirancang khusus untuk peringkasan teks abstraktif. Dikembangkan oleh Google Research dan dibangun di atas arsitektur Transformer, PEGASUS mengadopsi strategi pra-pelatihan yang meniru tugas peringkasan pada tahap *downstream*. Selama pra-pelatihan, model menghapus kalimat-kalimat yang dianggap paling *salient* dari dokumen sumber dan dilatih untuk merekonstruksi kembali kalimat-kalimat tersebut seolah-olah menjadi ringkasan dokumen. Objektif GSG menyelaraskan pra-pelatihan dengan tujuan peringkasan sehingga meningkatkan performa ke tugas *downstream*. Secara empiris, PEGASUS menunjukkan performa unggul dibandingkan dengan model berbasis Transformer lainnya pada *benchmark* peringkasan abstraktif [6], [9], [15].

Gambar 1 menyajikan gambaran umum arsitektur PEGASUS beserta dua objektif pra-pelatihan, *Gap Sentence Generation* (GSG) dan *Masked Language Modeling* (MLM). PEGASUS mengikuti desain *encoder–decoder* Transformer konvensional, di mana *encoder* menerima input yang telah di-*masking* dan menghasilkan representasi kontekstual yang digunakan *decoder* untuk merekonstruksi bagian yang hilang. Selama pra-pelatihan, seluruh kalimat yang dianggap *salient* digantikan oleh token *mask* khusus untuk objektif GSG, sedangkan token tambahan di-*masking* pada tingkat token untuk objektif MLM. Sebagai contoh, kalimat "Pegasus is mythical. It is pure white. It names the model." dapat diubah menjadi "Pegasus is [MASK2]. [MASK1] It [MASK2] the model." Dalam skema ini, token [MASK1] menandai penghapusan sebuah kalimat untuk GSG dan token [MASK2] menandai *masking* lokal untuk MLM. *Decoder* dilatih dengan *teacher forcing* dan menerima versi target yang digeser ke kanan (*shifted-right*) sebagai input awal, sehingga memungkinkan model mempelajari generasi autoregresif untuk membangun kembali teks yang hilang.



Gambar 1. Arsitektur PEGASUS

2.2. Tuning Bayesian

Tuning Bayesian adalah suatu metode probabilistik iteratif untuk secara efisien mengidentifikasi konfigurasi *hyperparameter* yang berkinerja tinggi bagi model pembelajaran mesin [16]. Berbeda dengan *grid search* atau *random search*, *tuning* Bayesian membangun model *surrogate* probabilistik, umumnya *Gaussian process*, untuk mengaproksimasi fungsi objektif yang mahal dan mengarahkan evaluasi berikutnya ke wilayah dengan ekspektasi peningkatan tinggi, sehingga mengurangi jumlah eksperimen mahal yang diperlukan untuk model berskala besar seperti arsitektur berbasis Transformer. Fungsi akuisisi, misalnya *Expected Improvement* atau *Upper Confidence Bound*, mengatur kompromi antara eksplorasi dan eksploitasi ketika memilih pengaturan kandidat *hyperparameter* untuk dievaluasi. Perpustakaan kontemporer seperti Optuna dan Hyperopt mengimplementasikan metodologi ini dan menjadikan pencarian *hyperparameter* yang otomatis serta hemat sampel praktis untuk alur kerja pembelajaran mendalam.

2.3. ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) adalah rangkaian metrik berbasis kemiripan (*overlap*) referensi yang banyak digunakan untuk menilai kualitas ringkasan otomatis dengan membandingkannya terhadap ringkasan buatan manusia [17]. ROUGE mencakup beberapa varian, terutama ROUGE-N yang mengukur overlap n-gram dan ROUGE-L yang mengukur kesamaan berdasarkan panjang *longest common subsequence* (LCS). Dalam praktik, ROUGE-1 dan ROUGE-2 sering dilaporkan karena keduanya efektif menangkap kebermaknaan dan kesesuaian terhadap teks referensi [9]. Penelitian ini menggunakan varian F1 dari ROUGE-1, ROUGE-2, ROUGE-L, dan ROUGE-LSUM untuk mengkuantifikasi kesesuaian leksikal dan struktural antara keluaran model dan ringkasan referensi. Rumus untuk menghitung ROUGE-N, ROUGE-L, dan ROUGE-LSUM ditunjukkan pada Persamaan 1, 2, dan 3.

$$\text{ROUGE-N} = \frac{\text{m}}{\text{n}} \tag{1}$$

Pada Persamaan 1, m menyatakan jumlah *n*-gram pada ringkasan kandidat yang tumpang tindih dengan *n*-gram pada ringkasan referensi, sedangkan n menyatakan jumlah total *n*-gram pada ringkasan referensi. Dengan kata lain, pembilang menghitung total n-gram referensi yang berhasil dipulihkan oleh kandidat (dengan pemangkasan sesuai hitungan referensi), dan penyebut adalah jumlah total *n*-gram dalam referensi.

$$\text{ROUGE-L} = \frac{\text{LCS}(\text{R}, \text{C})}{|\text{R}|} \tag{2}$$

Pada Persamaan 2, $\text{LCS}(\text{R}, \text{C})$ adalah panjang *longest common subsequence* antara referensi R dan kandidat C, dan $|\text{R}|$ adalah panjang referensi dalam satuan token. ROUGE-L mengekspresikan fraksi dari urutan referensi yang tercakup oleh subsekuensi cocok terpanjang tersebut.

$$\text{ROUGE-LSUM} = \frac{\text{LCS}_{\text{sum}}(\text{R}, \text{C})}{\sum_{\text{s}\in\text{R}}|\text{s}|} \tag{3}$$

Pada Persamaan 3, $\text{LCS}_{\text{sum}}(\text{R}, \text{C})$ menyatakan panjang agregat LCS yang dihitung melintasi pasangan kalimat yang dipasang antara referensi dan kandidat, dan penyebut $\sum_{\text{s}\in\text{R}}|\text{s}|$ adalah jumlah total token pada ringkasan referensi. ROUGE-LSUM adalah varian tingkat-ringkasan dari ROUGE-L yang menormalkan kecocokan LCS agregat terhadap panjang total referensi.

2.4. BERTScore

BERTScore adalah metrik evaluasi berbasis model yang mengukur kesamaan semantik antara teks referensi dan teks yang dihasilkan model dengan menghitung kemiripan kosinus antara *embedding* token kontekstual yang diekstrak dari *encoder* Transformer pra-latih [18]. Dengan menyelaraskan token berdasarkan kemiripan *embedding* dan mengagregasi penyelarasan tersebut menjadi skor precision, recall, dan F1, BERTScore menangkap korespondensi semantik yang sering terlewat oleh metrik n-gram literal seperti ROUGE atau BLEU, sehingga sangat sesuai untuk tugas di mana pelestarian makna lebih penting daripada kesamaan permukaan, misalnya, peringkasan abstraktif. Studi empiris menunjukkan bahwa BERTScore sering memberikan penilaian yang lebih representatif terhadap kesamaan kontekstual dan semantik dibandingkan dengan ukuran leksikal murni [18]. Penelitian ini melaporkan varian F1 dari BERTScore, yaitu rata-rata harmonis dari precision dan recall. Formulasi Precision, Recall, dan F1 masing-masing ditunjukkan pada Persamaan 4, 5, dan 6.

$$\text{Precision} = \frac{1}{|\text{C}|}\sum_{\text{x}\in\text{C}} \max_{\text{y}\in\text{R}} \text{sim}\,(\text{x}, \text{y}) \tag{4}$$

$$\text{Recall} = \frac{1}{|R|}\sum_{y\in R} \max_{x\in C} \text{sim}\,(x, y) \tag{5}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Pada Persamaan 4, C menyatakan kumpulan *embedding* token yang diekstrak dari teks kandidat (hasil model). Pada Persamaan 5, R menyatakan kumpulan *embedding* token yang diekstrak dari teks referensi (target). Fungsi $\text{sim}(\text{x}, \text{y})$ adalah kemiripan kosinus antara dua *embedding* token *x* dan *y*.

3. Metodologi Penelitian

3.1 Data Penelitian

Proses pengumpulan dan pembersihan data dimulai dengan konstruksi korpus berita berbahasa Indonesia yang diambil dari JPNN.com untuk periode Januari 2021 hingga Desember 2023. Dari total 360.026 artikel yang dikumpulkan, tahap pembersihan mengeliminasi 73.749 entri sehingga korpus akhir berjumlah 286.277 artikel. Tahap pembersihan tersusun dalam tiga

tingkatan. Tingkat pertama adalah pembersihan dasar yang menanggulangi isu struktural seperti nilai hilang dan duplikat. Tingkat kedua adalah pembersihan teks yang menghapus artefak tekstual seperti tag HTML serta menormalkan tanda baca dan kapitalisasi. Tingkat ketiga adalah pembersihan pasca-teks yang dilakukan setelah teks dinyatakan bebas *noise*, meliputi penghapusan ringkasan ekstraktif, penghapusan *string* kosong, dan penyaringan artikel yang terlalu pendek. *Dataset* akhir memuat empat kolom: badan artikel, judul, ringkasan, dan kata-kunci. Statistik ringkasan untuk korpus yang dibersihkan disajikan pada Tabel 1.

Tabel 1. Statistik Dataset

Varian	Jumlah Dokumen			Persentase N-gram Baru pada Ringkasan (%)			
	Latih	Validasi	Uji	1-gram	2-gram	3-gram	4-gram
Canon	257,649	14,314	14,314	16.7	44.9	61.1	70.4
Xtreme	257,649	5,595	5,595	26.6	67.9	88.6	97.6

Setelah pembersihan, tahap *preprocessing* menyiapkan data menjadi input model yang seragam. Pipeline diawali dengan pengacakan (*shuffling*) untuk menghilangkan ketergantungan temporal, lalu pemisahan awal menjadi data latih dan data validasi/tes dengan rasio 90/10. Bagian validasi/tes kemudian dibagi 50/50 menjadi validasi dan tes. Dari pembagian ini dibentuk dua varian *dataset*: Canon, yang mempertahankan keseluruhan sampel validasi dan tes, serta Xtreme, yang disusun dengan memfilter sampel validasi/tes untuk mempertahankan hanya item dengan tingkat kebaruan 4-gram pada kolom ringkasan yang melebihi 90%, sehingga menghasilkan set evaluasi yang lebih abstrak dan menantang. Masing-masing varian diproses melalui dua skema *preprocessing*: Reguler dan Informatif. *Preprocessing* Informatif menambahkan konteks dari judul dan kata kunci ke dalam badan artikel dan meregistrasikan beberapa token khusus pada tokenizer. Selain augmentasi ini, kedua skema melakukan tokenisasi, pemangkasan (truncation), dan konversi teks menjadi token ID yang sama. Contoh input reguler dan informatif ditampilkan pada Tabel 2.

Tabel 2. Contoh Input Reguler dan Informatif

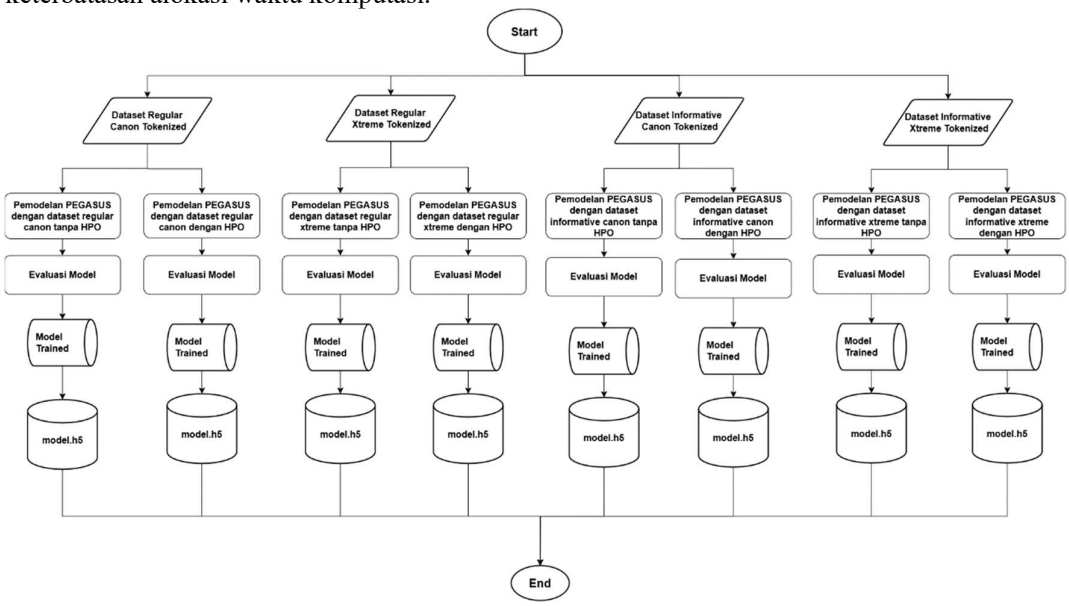
Input Reguler	Input Informatif
, jakarta lob 2 and pr group manager pt. bintang toedjoe andry mahyudi menyatakan pergantian musim membuat daya tahan tubuh kerap kali drop. bmkg menjelaskan bediding melanda sebagian selatan indonesia yang ... laporan bediding banyak dirasakan di wilayah tersebut. diperkirakan, bediding akan dialami hingga september 2025 cuaca yang bisa berubah drastis karena pancaroba serta bediding membawa sejumlah risiko kesehatan, terutama pada pernapasan. karena kelembapan udara yang rendah, tubuh jadi lebih rentan iritasi, terutama bagi mereka yang jarang bergerak aktif atau memiliki daya tahan tubuh rendah.	<TITLE> komix herbal hadirkan potek dance fest dalan kampanye edukasi kesehatan <KEYPHRASES> komix herbal <sep> bmkg <sep> batuk <sep> cuaca <sep> kampanye <sep> jakarta <ARTICLE> , jakarta lob 2 and pr group manager pt. bintang toedjoe andry mahyudi menyatakan pergantian musim membuat daya tahan tubuh kerap kali drop. bmkg menjelaskan bediding melanda sebagian selatan indonesia yang ... laporan bediding banyak dirasakan di wilayah tersebut. diperkirakan, bediding akan dialami hingga september 2025 cuaca yang bisa berubah drastis karena pancaroba serta bediding membawa sejumlah risiko kesehatan, terutama pada pernapasan. karena kelembapan udara yang rendah, tubuh jadi lebih rentan iritasi, terutama bagi mereka yang jarang bergerak aktif atau memiliki daya tahan tubuh rendah.

3.2 Alur Pemodelan

Alur pemodelan mencakup *fine-tuning*, evaluasi, dan penerapan HPO pada beberapa skenario, sebagaimana dijabarkan pada Gambar 2. Model PEGASUS yang telah dipra-latih pada korpus berbahasa Indonesia (thonyyy/pegasus_indonesian_base-finetune) di-*fine-tune* pada empat varian *dataset* yang telah ditokenisasi, yaitu Regular Canon, Regular Xtreme, Informative Canon, dan Informative Xtreme. Masing-masing varian dilatih dalam dua rezim, dengan dan tanpa *hyperparameter optimization* (HPO), sehingga menghasilkan delapan skenario eksperimen. Kinerja dinilai menggunakan ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSUM, dan

BERTScore, kemudian model final disimpan dalam format HDF5 (.h5). *Notebook* dan *dataset* yang digunakan pada penelitian juga dipublikasikan secara terbuka pada repositori GitHub[1].

Parameter tetap pelatihan mencakup ukuran *batch* 32, *epoch* maksimum 20, *optimizer* AdamWeightDecay, dan *random seed* 42, dengan *early stopping* yang memantau *validation loss* menggunakan *patience* 3 *epoch*. Seluruh eksperimen dijalankan pada perangkat keras TPU v4-16. Untuk skenario HPO, pencarian dilakukan dengan *tuning* Bayesian melalui perpustakaan Optuna dengan 12 *trial*, dibatasi pada tiga *hyperparameter* utama yakni *learning rate* (5×10^{-6} hingga 5×10^{-5}, log-uniform), *weight decay* (10^{-2} hingga 10^{-1}, log-uniform), dan *dropout rate* (0,1 hingga 0,5, uniform). Skenario tanpa HPO menggunakan nilai *default* umum untuk *fine-tuning* PEGASUS. Jumlah *trial* dan cakupan ruang pencarian yang relatif kecil merupakan konsekuensi keterbatasan alokasi waktu komputasi.



Gambar 2. Alur Pemodelan

4. Hasil dan Diskusi

Delapan skenario pemodelan dijalankan: Regular Canon, Regular Canon Tuned, Regular Xtreme, Regular Xtreme Tuned, Informative Canon, Informative Canon Tuned, Informative Xtreme, dan Informative Xtreme Tuned. Skenario tersebut dikarakterisasi menurut tiga dimensi eksperimen: skema input (ditambahkan atau tidaknya konteks tambahan), skema *dataset* (penggunaan varian *dataset* Canon atau Xtreme), dan skema *tuning* (diterapkan atau tidaknya HPO). Tabel 3 merangkum hasil evaluasi untuk setiap skenario.

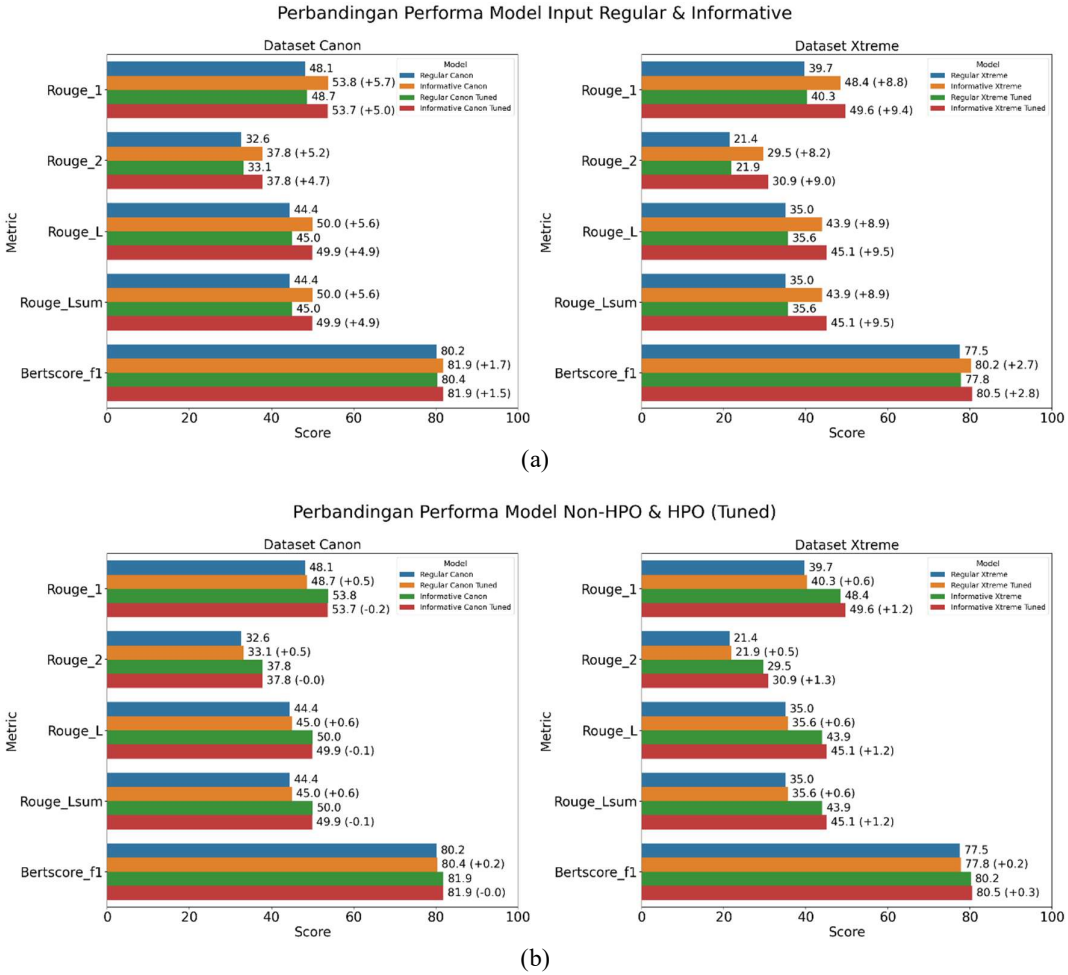
Tabel 3. Hasil Evaluasi Pemodelan

Model	Rouge_1	Rouge_2	Rouge_L	Rouge_Lsum	BERTScore_F1
Regular Canon	48,15	32,62	44,43	44,43	80,21
Regular Canon Tuned	48,70	33,15	44,99	44,99	80,40
Regular Xtreme	39,67	21,36	35,02	35,02	77,53
Regular Xtreme Tuned	40,28	21,88	35,63	35,63	77,77
Informative Canon	**53,82**	**37,82**	**50,01**	**50,01**	**81,91**
Informative Canon Tuned	53,67	37,81	49,92	49,92	81,88
Informative Xtreme	48,45	29,55	43,93	43,93	80,21
Informative Xtreme Tuned	**49,65**	**30,85**	**45,12**	**45,12**	**80,55**

Dalam eksperimen yang menggunakan *dataset* Canon, skenario Informative Canon memberikan hasil terbaik, mencapai skor tertinggi untuk ROUGE-1 (53,82%), ROUGE-2

[1] https://github.com/Arthdrm/Bayesian-Tuning-PEGASUS-2025

(37,82%), ROUGE-L (50,01%), dan ROUGE-LSUM (50,01%), serta BERTScore F1 sebesar 81,91%. Untuk *dataset* Xtreme, performa terbaik diamati pada skenario Informative Xtreme Tuned, yang memperoleh ROUGE-1 (49,65%), ROUGE-2 (30,85%), ROUGE-L (45,12%), ROUGE-LSUM (45,12%), dan BERTScore F1 sebesar 80,55%. Label "Tuned" menunjukkan penerapan *hyperparameter* optimization (HPO) selama proses *fine-tuning*.



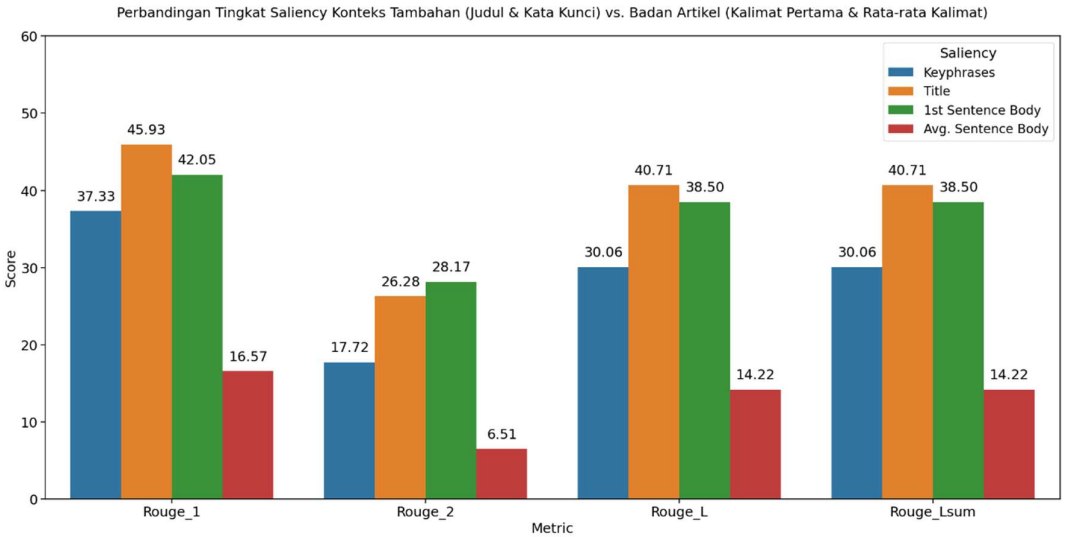
Gambar 3. Perbandingan Performa Model
(a) Skema Input Reguler vs Informatif (b) Skema Tuning Non-HPO vs HPO

Gambar 3(a) menunjukkan bahwa untuk kedua skema *dataset*, baik *dataset* Canon maupun Xtreme, model yang dilatih dengan konteks tambahan pada input memiliki kinerja lebih baik dibandingkan dengan model yang dilatih dengan input reguler. Pada skema Canon, varian Informative menghasilkan peningkatan rata-rata sebesar +5,32 poin (+10,9%) untuk ROUGE-1, +4,93 poin (+14,9%) untuk ROUGE-2, +5,25 poin (+11,7%) untuk ROUGE-L, +5,25 poin (+11,7%) untuk ROUGE-LSUM, dan +1,59 poin (+1,9%) untuk BERTScore F1. Pada skema Xtreme, peningkatan yang diamati lebih besar: +9,07 poin (+22,6%) untuk ROUGE-1, +8,58 poin (+39,6%) untuk ROUGE-2, +9,20 poin (+26,0%) untuk ROUGE-L, +9,20 poin (+26,0%) untuk ROUGE-LSUM, dan +2,70 poin (+3,5%) untuk BERTScore F1. Temuan ini sejalan dengan hasil yang dilaporkan oleh [12], yang juga menemukan kenaikan performa setelah penambahan konteks untuk tugas peringkasan abstraktif berbahasa Indonesia. Namun, perlu dicatat bahwa perbandingan angka tidak langsung dapat dilakukan karena perbedaan model (IndoBERT,

mBART, mT5 pada studi tersebut) dan perbedaan korpus (Liputan6). Sebagai pembanding tambahan, studi peringkasan berita berbahasa Indonesia menggunakan mBART pada korpus XL-Sum ID melaporkan skor ROUGE-1 sebesar 35,94%, ROUGE-2 16,43%, dan ROUGE-L 29,91% [19]. Skor ROUGE dengan rata-rata 50 poin yang diperoleh skenario *Informative Canon* menempatkan pendekatan *input* informatif di atas spektrum performa yang umumnya dilaporkan untuk korpus berita berbahasa Indonesia, meskipun perbandingan langsung tetap dibatasi oleh perbedaan korpus, *preprocessing*, dan konfigurasi evaluasi.

Gambar 3(b) menunjukkan bahwa penerapan HPO, khususnya *tuning* Bayesian, tidak selalu menghasilkan peningkatan performa yang konsisten. Dari empat skenario pemodelan yang menerapkan HPO, tiga menunjukkan kenaikan performa, sementara satu mengalami penurunan. Penurunan terjadi pada skenario Informative Canon Tuned, di mana skor model hasil *tuning* turun sebesar −0,15 poin (−0,27%) untuk ROUGE-1, −0,01 poin (−0,02%) untuk ROUGE-2, −0,09 poin (−0,17%) untuk ROUGE-L, −0,09 poin (−0,17%) untuk ROUGE-LSUM, dan −0,03 poin (−0,03%) untuk BERTScore F1 jika dibandingkan dengan Informative Canon. Selain itu, peningkatan yang diamati pada tiga skenario yang membaik bersifat terbatas. Rata-rata kenaikan di seluruh metrik evaluasi tidak melebihi tiga persen. Temuan ini tidak sepenuhnya selaras dengan [14], yang melaporkan efek positif *tuning* Bayesian pada model Transformer untuk peringkasan.

Peningkatan konsisten pada skema *input* informatif dapat dijelaskan secara mekanistik. Judul dan kata kunci di awal input berfungsi sebagai sinyal *salient* terkompresi yang selaras dengan objektif pra-pelatihan *Gap Sentence Generation* PEGASUS, sementara penggunaan token khusus <TITLE>, <KEYPHRASES>, dan <ARTICLE> memberikan penandaan struktural eksplisit sehingga model tidak perlu menginferensi peran dari setiap segmen input. Interpretasi ini konsisten dengan analisis *saliency* pada Gambar 4. Sebaliknya, efek marginal dari Bayesian *hyperparameter tuning* dapat disebabkan oleh beberapa faktor. Faktor pertama adalah bahwa model pralatih sudah mendekati wilayah *loss* yang baik, sehingga *fine-tuning* mengalami *diminishing returns*. Faktor selanjutnya adalah keterbatasan jumlah *trial* dan pilihan *hyperparameter* yang membatasi eksplorasi optimal, serta perbedaan model dasar (T5 versus PEGASUS) dan skala eksperimen yang menjelaskan disparitas temuan dibandingkan dengan [14].



Gambar 4. Perbandingan Tingkat *Saliency* Konteks Tambahan vs Badan Artikel

Gambar 4 menyajikan analisis perbandingan tingkat *saliency* (bobot kalimat) untuk informasi kontekstual tambahan pada skema input Informative (judul dan kata kunci) terhadap ukuran *saliency* yang dihitung dari badan artikel (kalimat pertama dan rata-rata kalimat). *Saliency*

dikuantifikasi menggunakan varian ROUGE (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSUM) sebagai proksi kepentingan, mengikuti praktik penggunaan ROUGE-1 pada pra-pelatihan PEGASUS dengan objektif GSG. Semua pengukuran dihitung terhadap kolom ringkasan pada partisi pelatihan (N = 257.649). Hasil menunjukkan bahwa judul mencapai tingkat *saliency* yang lebih tinggi dibandingkan dengan kalimat pertama pada badan artikel pada hampir semua metrik ROUGE, dengan pengecualian ROUGE-2. Meskipun kata-kunci menunjukkan *saliency* yang lebih rendah daripada kalimat pertama, nilainya tetap melebihi *saliency* rata-rata kalimat badan artikel lebih dari dua kali lipat. Temuan ini mengindikasikan bahwa penambahan judul dan kata kunci ke dalam badan artikel mengonsentrasikan sinyal *salient* yang dapat membantu model mengidentifikasi kata dan kalimat kunci sebagai referensi dalam proses generasi ringkasan abstraktif.

5. Kesimpulan dan Saran

Performa model PEGASUS yang dilatih dengan input informatif, yaitu input yang diperkaya dengan konteks tambahan berupa judul berita dan kata kunci secara substansial, mengungguli model PEGASUS yang dilatih hanya dengan input reguler (tanpa konteks tambahan). Rata-rata peningkatan performa yang diamati ialah +16,75% untuk ROUGE-1, +27,25% untuk ROUGE-2, +18,58% untuk ROUGE-L, +18,58% untuk ROUGE-LSUM, dan +2,7% untuk BERTScore F1 dibandingkan model yang hanya menggunakan input reguler. Analisis *saliency* (bobot kalimat) terhadap konteks tambahan menunjukkan tingkat *saliency* yang konsisten tinggi untuk judul dan kata kunci.

Penerapan HPO, khususnya tuning Bayesian, terhadap model PEGASUS hanya menghasilkan kenaikan performa yang bersifat marginal. Rata-rata peningkatan yang tercatat adalah +1,21% untuk ROUGE-1, +2,10% untuk ROUGE-2, +1,38% untuk ROUGE-L, +1,38% untuk ROUGE-LSUM, dan +0,23% untuk BERTScore F1 dibandingkan dengan model tanpa tuning Bayesian. Hasil ini diperoleh dari jumlah *trial* dan ruang *hyperparameter* yang relatif kecil, yaitu sebesar 12 *trial* dan tiga *hyperparameter,* akibat keterbatasan waktu dan sumber daya penulis. Oleh karena itu, kemungkinan peningkatan yang lebih substansial dapat dicapai dengan jumlah *trial* dan ruang pencarian *hyperparameter* yang lebih besar.

Referensi

[1] W. Zarman, "Information Overload: Clarifying the Problem," *Indonesian Journal of Informatics Education*, vol. 5, no. 2, pp. 1–5, 2021, doi: 10.20961/ijie.v5i2.56922.

[2] A. M. F. Yousef, A. Alshamy, A. Tlili, and A. H. S. Metwally, "Demystifying the New Dilemma of Brain Rot in the Digital Era: A Review," *Brain Sciences*, vol. 15, no. 3, p. 283, 2025, doi: 10.3390/brainsci15030283.

[3] A. Eko Raharjo, "Profiling News Consumption on Social Media," *Jurnal Komunikasi Profesional*, vol. 5, no. 4, pp. 320–334, 2021, doi: 10.25139/jkp.v5i4.3794.

[4] D. Yadav, J. Desai, and A. K. Yadav, "Automatic Text Summarization Methods: A Comprehensive Review," *SN Computer Science*, vol. 4, no. 1, p. 33, 2022, doi: 10.1007/s42979-022-01446-w.

[5] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A Survey of Text Summarization: Techniques, Evaluation and Challenges," *Natural Language Processing Journal*, vol. 7, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.

[6] H. Lucky and D. Suhartono, "Investigation of Pre-Trained Bidirectional Encoder Representations from Transformers Checkpoints for Indonesian Abstractive Text Summarization," *Journal of Information and Communication Technology*, vol. 21, no. 1, pp. 71–94, 2022, doi: 10.32890/jict2022.21.1.4.

[7] Z. Alami Merrouni, B. Frikh, and B. Ouhbi, "EXABSUM: A New Text Summarization Approach for Generating Extractive and Abstractive Summaries," *Journal of Big Data*, vol. 10, no. 1, p. 163, 2023, doi: 10.1186/s40537-023-00836-y.

[8] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural Language Processing with Transformers: A Review," *PeerJ Computer Science*, vol. 10, p. e2222, 2024, doi: 10.7717/peerj-cs.2222.

[9] F. F. Kartamanah, A. R. Atmadja, and I. Budiman, "Analyzing PEGASUS Model Performance with ROUGE on Indonesian News Summarization," *Jurnal dan Penelitian Teknik Informatika*, vol. 9, no. 1, pp. 31–42, 2025, doi: 10.33395/sinkron.v9i1.14278.

[10] R. Alsultan et al., "PEGASUS-XL with Saliency-Guided Scoring and Long-Input Encoding for Multi-Document Abstractive Summarization," *Scientific Reports*, vol. 15, no. 1, p. 26529, 2025, doi: 10.1038/s41598-025-11062-2.

[11] A. Bahari and K. E. Dewi, "Peringkasan Teks Otomatis Abstraktif Menggunakan Transformer pada Teks Bahasa Indonesia," *KOMPUTA: Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 1, pp. 83–91, 2024, doi: 10.34010/komputa.v13i1.11197.

[12] F. Koto, T. Baldwin, and J. H. Lau, "LipKey: A Large-Scale News Dataset for Absent Keyphrases Generation and Abstractive Summarization," in *Proc. 29th Int. Conf. Computational Linguistics (COLING)*, Gyeongju, Republic of Korea, 2022, pp. 3427–3437, doi: 10.18653/v1/2022.coling-1.303.

[13] B. Bischl *et al.*, "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.

[14] A. R. Lubis *et al.*, "Enhancing Text Summarization with a T5 Model and Bayesian Optimization," *Revue d'Intelligence Artificielle*, vol. 37, no. 5, pp. 1213–1219, 2023, doi: 10.18280/ria.370513.

[15] C. M. Muia, A. M. Oirere, and R. N. Ndung'u, "A Comparative Study of Transformer-Based Models for Text Summarization of News Articles," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 13, no. 2, pp. 37–43, 2024, doi: 10.30534/ijatcse/2024/011322024.

[16] A. H. Victoria and G. Maragatham, "Automatic Tuning of Hyperparameters Using Bayesian Optimization," *Evolving Systems*, vol. 12, no. 1, pp. 217–223, 2021, doi: 10.1007/s12530-020-09345-2.

[17] A. Dalal et al., "Text Summarization for Pharmaceutical Sciences Using Hierarchical Clustering with a Weighted Evaluation Methodology," *Scientific Reports*, vol. 14, no. 1, p. 20149, 2024, doi: 10.1038/s41598-024-70618-w.

[18] Junadhi, Agustin, L. Efrizoni, F. Okmayura, D. R. Habibie, and Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 885–896, 2025, doi: 10.47738/jads.v6i2.547.

[19] R. H. Astuti, M. Muljono, and S. Sutriawan, "Indonesian News Text Summarization Using MBART Algorithm," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 155–164, 2024, doi: 10.15294/sji.v11i1.49224.