Consumer Health Question Summarization Berbasis Transfer Learning untuk Bahasa Indonesia

Ananda Alvin Bernadian H., Ananda Rafi Prasetyo, Ahmad Alden, Brahmantya Fikri Setya P., Riska Lathifah

Abstract

Background: Pencarian informasi kesehatan secara online adalah praktik yang semakin umum dilakukan oleh konsumen untuk mendapatkan informasi medis yang cepat dan mudah diakses. Namun, masih banyak pertanyaan kesehatan yang diajukan terlalu panjang dan mengandung rincian yang tidak relevan, yang justru mempersulit *search engine* dalam mencari jawaban yang tepat.

Objective: Penelitian ini bertujuan untuk mengembangkan model *consumer health question summarization* berbasis *transfer learning* untuk bahasa Indonesia, dengan fokus pada pengembangan korpus baru, eksplorasi *pre-trained model*, dan analisis kesalahan untuk meningkatkan kualitas ringkasan pertanyaan kesehatan konsumen.

Methods: Tahapan metodologi dari penelitian ini dimulai dengan proses *pre-processing* data, yang mencakup pembersihan teks, normalisasi bahasa, tokenisasi dan filter kata umum, serta penghapusan *stopword*. Selanjutnya, membandingkan enam model *summarization*.. Evaluasi kinerja model dilakukan menggunakan metrik ROUGE-1, ROUGE-2, ROUGE-L, dan BLEU untuk mengukur kesamaan antara ringkasan hasil model dengan teks referensi. Setelah evaluasi awal, model **UnifiedQA-T5** dipilih sebagai model terbaik karena memberikan performa tertinggi. Tahap selanjutnya adalah fine-tuning pada model **UnifiedQA-T5** untuk mengoptimalkan kinerjanya. Evaluasi ulang dilakukan setelah *fine-tuning* untuk memastikan peningkatan performa model berdasarkan metrik yang sama

Results: Hasil eksperimen menunjukkan bahwa model mBART-L dan Seq2Seq berhasil merangkum pertanyaan dengan mempertahankan relevansi dan makna dari teks asli. Berikut adalah hasil performa untuk masing-masing model berdasarkan evaluasi dengan metrik ROUGE dan BLEU: **mBART-L** memiliki **ROUGE-1** sebesar 0.2974; **ROUGE-2** sebesar 0.1274; **ROUGE-L** sebesar 0.2508; dan **BLEU Score**: 0.0202. Sedangkan **Seq2Seq** memiliki **ROUGE-1** sebesar 0.2312; **ROUGE-L** sebesar 0.2087; dan **BLEU Score**: 0.6090. Hasil ini menunjukkan bahwa model **Seq2Seq** memberikan performa yang sedikit lebih baik dibandingkan mBART-L, dengan skor ROUGE dan BLEU yang lebih tinggi, menandakan kualitas ringkasan yang lebih relevan dan akurat dalam konteks pertanyaan kesehatan.

Conclusion: Penelitian ini berhasil mengembangkan model *summarization* berbasis transfer learning untuk merangkum pertanyaan panjang di platform *telemedicine* berbahasa Indonesia. Model UnifiedQA-T5, setelah *fine-tuning*, menunjukkan skor ROUGE-1 0.4992, ROUGE-2 0.2800, dan ROUGE-L 0.4760, dan BLEU 0.1676. Model ini meningkatkan efisiensi sistem *Question Answering (QA)* di platform seperti Alodokter, dengan hasil ringkasan yang lebih relevan dan akurat. Namun, masih ada ruang perbaikan untuk menangani pertanyaan yang lebih kompleks.

Keywords: Telemedicine, Natural Language Processing, Text Summarization, Model T5, Question Summarization.

Introduction

Dengan pesatnya kemajuan teknologi dan penggunaan internet, semakin banyak konsumen yang mencari informasi terkait kesehatan secara daring. Berdasarkan survei oleh Rock Health pada tahun 2023, lebih dari 80% orang di seluruh dunia telah mengakses layanan kesehatan melalui *telemedicine* setidaknya sekali dalam hidup mereka (sitasi rock health). Fenomena ini menunjukkan betapa pentingnya platform digital dalam menyediakan layanan kesehatan yang lebih mudah diakses oleh masyarakat. Namun, meskipun kemudahan ini sangat menguntungkan, banyak pertanyaan yang diajukan oleh pasien atau konsumen dalam mencari informasi kesehatan masih terlalu panjang dan mengandung rincian yang tidak relevan (Shweta Yadav, towards). Hal ini membuat pencarian jawaban yang tepat menjadi lebih sulit, baik bagi pengguna maupun *search engine* (Ghafouri Arash).

Salah satu pendekatan untuk mengatasi masalah ini adalah melalui *Question Summarization* (QS), yang bertujuan untuk merangkum pertanyaan yang panjang dan kompleks menjadi bentuk yang lebih ringkas dan relevan tanpa mengurangi makna inti pertanyaan (balumuri, spandana). Proses ini sangat penting dalam konteks pertanyaan kesehatan konsumen yang sering kali mengandung informasi latar belakang yang tidak relevan, seperti riwayat medis pasien (sweta yadav nlm at medica). Mengurangi atau menghilangkan informasi yang tidak perlu dalam pertanyaan ini dapat membantu sistem *question answering* (QA) dalam memberikan jawaban yang lebih tepat dan efisien (nour eddine, enhancing large).

Question:Dok saya pakai softlens tiap hari udah lama sejak 6 tahun yg lalu, aman ama aja tidak ada kendala dok,, biar ga kering bgt mata nya pake obat tetes apa ya dok bagusnya?

Summary: Apa obat tetes untuk mata kering?

Gambar 1. Pertanyaan ini mencakup detail gejala fisik pada bayi yang mengkhawatirkan. Ringkasan menyederhanakan informasi menjadi inti masalah untuk mempermudah analisis.

Question:Dok mau bertanya,,, saya rutin olahraga lari sore dan menghitung aktifitas saya dg smartwatch, apakah itu valid? Dan bagaimana cara menghitung denyut nadi yg benar dan valid dok? Mohon petunjuknya dok, thanks dokter..

Summary: Bagaimana cara menghitung denyut nadi?

Gambar 2. Fokus pada kebutuhan solusi alami (jamu) untuk meningkatkan berat badan. Ringkasan menyoroti inti masalah secara padat.

Seiring dengan perkembangan dalam pemrosesan bahasa alami (NLP), berbagai model pra-latih, seperti PEGASUS (*Pre-Training with Extracted Gap-Sentences for Abstractive Summarization*) (lung hao lee, NCUEE), BART, dan Mini LM, telah digunakan untuk tugas ringkasan teks, termasuk dalam domain medis (sweta yadav, question aware). Meskipun hasil yang dicapai dalam bahasa Inggris sangat menjanjikan, tantangan besar tetap ada dalam penerapan teknik ini untuk bahasa selain bahasa Inggris, seperti bahasa Indonesia. Perbedaan struktural antara kedua bahasa, seperti fleksibilitas urutan kata dalam bahasa Indonesia yang lebih bergantung pada konteks, mempersulit model dalam memproses dan mendeteksi informasi yang relevan. Selain itu, variasi kosakata, dialek, dan penggunaan istilah medis yang tidak terstandarisasi memperburuk pemrosesan teks dalam konteks kesehatan. Keterbatasan korpus berbahasa Indonesia yang telah dianotasi dengan baik juga menjadi hambatan, karena bahasa Indonesia belum memiliki dataset besar dan berkualitas tinggi seperti bahasa Inggris. Kurangnya data yang relevan serta model *pre-trained* yang efektif untuk bahasa Indonesia membuat penerapan teknik ini lebih kompleks, memerlukan penyesuaian model dan pendekatan yang hati-hati agar dapat menghasilkan ringkasan yang akurat dan relevan (rini wijayanti).

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model *consumer health question summarization* berbasis *transfer learning* untuk bahasa Indonesia. Pendekatan ini diharapkan dapat merangkum pertanyaan kesehatan konsumen dengan lebih efektif, dengan mempertimbangkan karakteristik dan tantangan bahasa Indonesia yang berbeda dari bahasa Inggris. Kontribusi utama dari penelitian ini adalah sebagai berikut: **(i)** mendefinisikan ulang *Question Summarization* sebagai proses untuk menghasilkan pertanyaan yang terkompresi yang hanya mencakup informasi minimum yang diperlukan untuk menemukan jawaban yang tepat,

dan mengembangkan korpus baru dari pertanyaan kesehatan konsumen berbahasa Indonesia beserta ringkasannya (lihat gambar 1 dan 2); **(ii)** mengeksplorasi penggunaan model *pretrained* atau *transfer learning* dalam konteks *question summarization* untuk bahasa Indonesia, serta menganalisis kinerjanya pada dataset pertanyaan kesehatan konsumen berbahasa Indonesia; **(iii)** melakukan analisis kesalahan secara rinci dan membahas potensi perbaikan dalam tugas ringkasan pertanyaan kesehatan konsumen berbahasa Indonesia.

Related Works

Dalam beberapa tahun terakhir, penelitian tentang *text summarization* telah mengalami perkembangan pesat, terutama untuk bahasa-bahasa dengan sumber daya melimpah, seperti bahasa Inggris. Sebagian besar penelitian ini fokus pada tugas peringkasan teks dalam domain seperti berita, artikel, atau ulasan produk, dengan memanfaatkan pendekatan ekstraktif dan abstraktif. Pendekatan ekstraktif bekerja dengan mengambil kalimat-kalimat penting dari dokumen asli, sementara pendekatan abstraktif berusaha menghasilkan ringkasan yang lebih menyerupai cara manusia merangkum teks melalui parafrasa. Berbagai penelitian juga telah menunjukkan efektivitas penggunaan *pre-trained models* seperti BERT dan GPT untuk meningkatkan akurasi peringkasan teks (Wijayanti et al., 2021). Sedangkan pendekatan abstraktif adalah metode yang bertujuan untuk menghasilkan ringkasan dengan menulis ulang atau memparafrase teks asli, sehingga ringkasan yang dihasilkan lebih menyerupai ringkasan buatan manusia (Wijayanti et al., 2021). Contohnya, metode BERT2GPT menunjukkan performa yang baik dalam *abstractive summarization* dengan nilai ROUGE-L sebesar 0.60, memadukan kemampuan representasi kontekstual dari BERT dan kemampuan generatif dari GPT-2 (Nasari et al., 2023).

Untuk bahasa Indonesia, dataset IndoSum menjadi salah satu *benchmark* yang banyak digunakan untuk penelitian peringkasan teks otomatis. IndoSum menyediakan lebih dari 19.000 pasangan artikel berita dan ringkasan, dan telah digunakan dalam penelitian abstraktif maupun ekstraktif. Penelitian dengan menggunakan BERT dan model pra-latih lainnya menunjukkan potensi besar dalam peringkasan teks berbahasa Indonesia, meskipun hasilnya masih kalah optimal dibanding bahasa Inggris (Wijayanti et al., 2021). Namun, meskipun banyak penelitian mengenai *text summarization* secara umum, studi mengenai *question summarization*, terutama untuk bahasa selain Inggris, masih sangat terbatas. Dalam konteks bahasa Indonesia, sebagian besar penelitian tetap berfokus pada peringkasan berita atau artikel, dengan hanya sedikit eksplorasi pada *question summarization.* Salah satu kendala utama adalah keterbatasan dataset spesifik untuk tugas ini, yang menghambat pengembangan model yang efektif (Saputra & Maki, 2021).

Penelitian pada *question summarization* dalam domain kesehatan telah dilakukan dalam bahasa Inggris. Sebagai contoh, Ben Abacha dan Demner-Fushman (2019) [32] memperkenalkan korpus MeQSum yang berisi 1.000 pertanyaan kesehatan konsumen beserta ringkasannya, yang menunjukkan bahwa pendekatan *pointer-generator network* mencapai skor ROUGE-1 sebesar 44,16% pada tugas ini. Selain itu, sistem NCUEE-NLP menggunakan model PEGASUS Transformer dalam MEDIQA 2021 untuk meringkas pertanyaan kesehatan konsumen dengan hasil yang signifikan, mencapai peringkat ketiga dalam kompetisi tersebut dengan skor ROUGE-2-F1 sebesar 0,1597 (Lee et al., 2021) [27].

Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan mengusulkan pendekatan baru untuk question summarization dalam bahasa Indonesia. Dengan adanya pengembangan ini, diharapkan dapat mendukung terciptanya sistem tanya jawab otomatis yang lebih efisien dan akurat.

Methods

Bagian ini menjelaskan secara rinci mengenai data yang digunakan serta metodologi yang diterapkan dalam penelitian. Penjelasan mencakup sumber, karakteristik, dan pengolahan data, diikuti dengan langkah-langkah sistematis dalam pelaksanaan penelitian untuk mencapai tujuan yang telah ditentukan.

1. Material

Dataset yang digunakan dalam penelitian ini didapatkan dari proses *scraping* dari laman diskusi Alodokter. Dataset ini terdiri dari 1000 baris data dengan terdiri dari 4 kolom utama, yaitu:

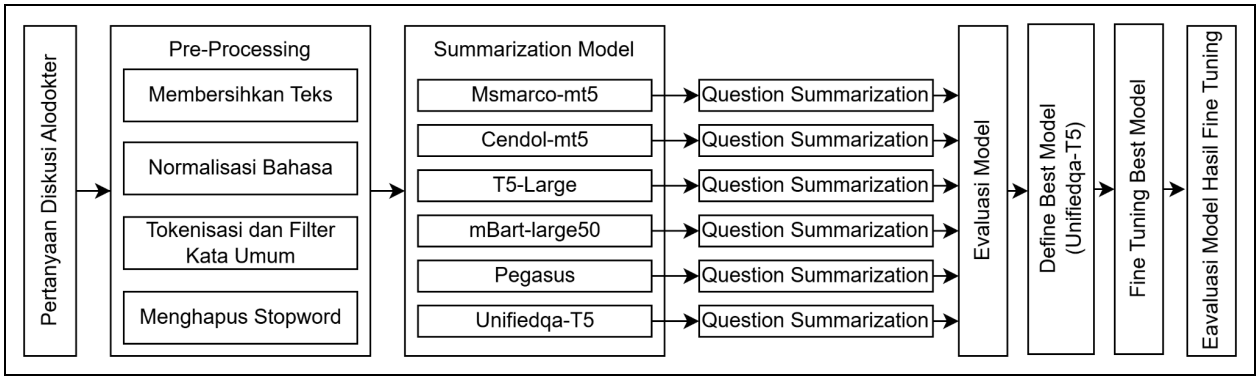
- **Nama**: Nama pengguna yang mengajukan pertanyaan (disamarkan untuk menjaga privasi).
- **Pertanyaan Singkat**: Ringkasan singkat dari pertanyaan yang diajukan.
- **URL Pertanyaan Lengkap**: Alamat URL yang mengarah ke laman detail pertanyaan di situs Alodokter.
- **Pertanyaan Panjang**: Teks lengkap dari pertanyaan yang diajukan oleh pengguna.

Proses *scraping* dilakukan dengan memperhatikan berbagai aspek, salah satunya adalah memastikan hanya data yang relevan dan terbuka untuk publik yang diambil, tanpa melanggar privasi pengguna. Dataset ini mencakup berbagai topik diskusi kesehatan, mulai dari keluhan umum hingga pertanyaan spesifik terkait gejala, diagnosa, atau cara pengobatan.

Dengan jumlah total 1000 data, dataset ini cukup besar untuk digunakan dalam berbagai jenis analisis berbasis Natural Language Processing (NLP) dengan menggunakan berbagai metode, seperti *text summarization.* Dataset ini merepresentasikan pola umum diskusi kesehatan di laman tersebut dan memberikan wawasan yang luas untuk penelitian.

2. Methodology

Penelitian ini melibatkan beberapa tahapan utama, dimulai dari pengumpulan data, preprocessing data, implementasi model, hingga evaluasi hasil. Setiap tahapan dirancang secara sistematis untuk menghasilkan solusi yang relevan terhadap permasalahan yang diangkat.



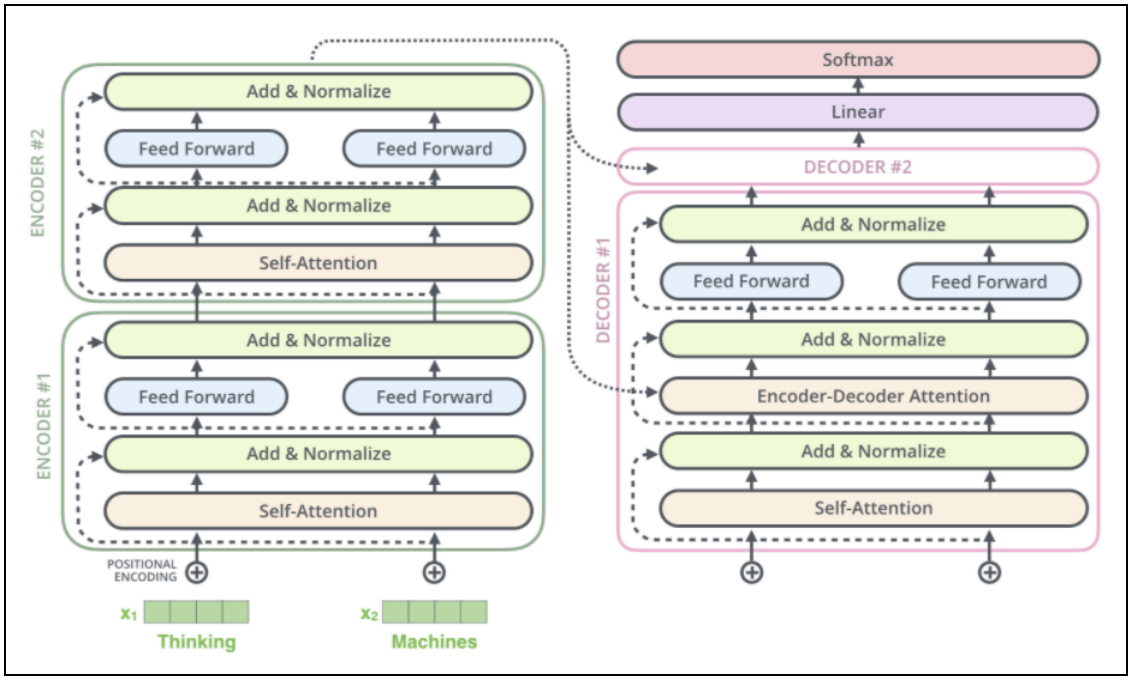
Gambar 3. Diagram Flowchart Tahapan Penelitian

Tahapan *preprocessing* data digunakan bertujuan untuk membersihkan data mentah dan untuk mempersiapkan agar dapat diproses oleh model *summarization.* Tahapan *preprocessing* yang digunakan mencakup beberapa langkah yaitu, yang pertama adalah membersihkan teks dari karakter yang tidak relevan, seperti spasi berlebih dan juga menghilangkan data pribadi sensitif seperti nama dan lokasi. Hal itu dilakukan untuk tetap menjaga privasi dari pengguna (Ben Abacha & Demner-Fushman, 2019; Zhang et al., 2022).

Langkah selanjutnya merupakan normalisasi bahasa, dimana kata atau istilah yang tidak baku diubah menjadi kata yang baku atau standar. Contohnya adalah seperti “sy” menjadi “saya” dan kata “gk” menjadi “tidak”. Normalisasi bahasa ini yang dilakukan menggunakan sebuah kamus singkatan yang dibuat secara manual berdasarkan analisis kata-kata dalam teks asli yang ada dalam dataset mentah (Lubis et al., 2023; Yadav et al., 2021).

Untuk memastikan data yang akan diproses hanya mengandung konteks yang diperlukan, maka dilakukan sebuah tokenisasi dan penghapusan kata umum dengan memisahkan kata-kata dan menghapus kata umum yang tidak diperlukan seperti “halo”, “assalamualaikum”, atau “terima kasih” karena tidak berpengaruh pada hasil pemodelan yang akan digunakan. Terakhir, melakukan penghapusan *stopword*, yaitu kata yang sering sekali muncul tetapi tidak memberikan sebuah makna yang signifikan dalam proses pemodelan. *Stopword* dihapus bertujuan untuk menyederhanakan teks yang akan diproses dalam model sehingga dapat mengefisiensikan fungsi dari model (Mrini et al., 2021; Widodo et al., 2021; Zhang et al., 2022).

Pada tahap *Question Summarization*, proyek ini terinspirasi oleh penelitian sebelumnya (Lubis et al., 2023; Yadav, Sarrouti, & Gupta, 2021) yang menggunakan model T5 untuk tugas *text summarization*, selain itu, terus dilakukan eksplorasi variasi dari model T5 ini. Salah satu model yang ditemukan dan dipilih untuk penelitian ini adalah UnifiedQA-T5. Model ini dipilih karena kemampuannya dalam menghasilkan ringkasan pertanyaan yang tetap ringkas dan relevan dengan teks aslinya tanpa memerlukan pelatihan ulang. UnifiedQA-T5 adalah model yang sebelumnya telah dilatih menggunakan dataset yang berisi pasangan pertanyaan dan jawaban, sehingga dapat digunakan untuk tugas *question summarization*, termasuk dalam meringkas pertanyaan panjang di domain kesehatan (Ben Abacha & Demner-Fushman, 2019; Zhang et al., 2022).



Gambar 4. Arsitektur Model T5 (Sumber: Wang, M. 2023)

Model UnifiedQA-T5 bekerja dengan pendekatan *encoder-decoder*, di mana encoder berfungsi untuk mengubah teks panjang menjadi representasi vektor yang bermakna, sementara decoder menghasilkan teks ringkasan berdasarkan representasi tersebut. Metode UnifiedQA-T5 dipilih karena model ini sudah memiliki kemampuan untuk memproses teks dalam berbagai bahasa, termasuk bahasa Indonesia, tanpa perlu pengolahan atau pelatihan tambahan. Hal ini sangat sesuai dengan keterbatasan sumber daya yang ada, sebagaimana dibahas dalam penelitian sebelumnya (Widodo et al., 2021; Mrini et al., 2021). Penemuan mengenai model ini berawal dari riset yang dilakukan di Hugging Face, di mana paper yang terkait dengan model ini, yaitu "UNIFIEDQA: Crossing Format Boundaries with a Single QA System" yang dipublikasikan oleh Khashabi et al. (2020).

Untuk mengevaluasi kualitas ringkasan yang dihasilkan oleh model, digunakan dua metrik utama, yaitu ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) dan BLEU (*Bilingual Evaluation Understudy*). Kedua metrik ini mengukur kesamaan antara ringkasan yang dihasilkan dan teks referensi, dengan fokus pada berbagai aspek kesamaan leksikal dan struktur teks. ROUGE digunakan untuk mengevaluasi tingkat kesamaan antara ringkasan hasil model dengan teks referensi. Metrik ini mengukur kesamaan berdasarkan berbagai tingkatan overlap kata-kata dalam teks, yang terdiri dari ROUGE-1, ROUGE-2, dan ROUGE-L[2][3].

ROUGE-1 berfungsi mengukur overlap kata tunggal antara teks referensi dan hasil ringkasan. Rumus ROUGE-1 adalah sebagai berikut:

$$ROUGE-1 = \frac{\sum_{n=1}^{N} \text{count of matched unigrams}}{\sum_{n=1}^{N} \text{count of unigrams in reference}}$$

Skor ini menunjukkan seberapa baik model menangkap kata-kata kunci yang ada dalam teks referensi. ROUGE-2 mengukur overlap pasangan kata (*bigrams*), yang penting untuk memastikan bahwa model dapat menangkap kombinasi kata yang relevan dalam konteks teks. Rumus ROUGE-2 adalah:

$$ROUGE-2 = \frac{\sum_{n=1}^{N} \text{count of matched bigrams}}{\sum_{n=1}^{N} \text{count of bigrams in reference}}$$

Skor ini menunjukkan sejauh mana model mempertahankan pasangan kata penting dalam teks. ROUGE-L mengukur kesamaan berdasarkan urutan kata terpanjang yang sama antara teks asli dan ringkasan yang dihasilkan. Rumus untuk menghitung ROUGE-L adalah:

$$ROUGE-L = \frac{\sum_{n=1}^{N} \text{length of longest common subsequence}}{\sum_{n=1}^{N} \text{length of subsequence in reference}}$$

ROUGE-L penting karena membantu memastikan bahwa inti dan makna dari teks asli tetap terjaga dalam ringkasan[1][2][3].

Sementara itu, BLEU adalah matrik evaluasi yang digunakan untuk mengukur seberapa mirip ringkasan hasil model dengan teks aslinya, dengan fokus pada kesamaan berbasis n-gram. BLEU mengukur dua aspek utama, yaitu kesesuaian struktur dan relevansi kata kunci. Kesesuaian Struktur mengukur sejauh mana teks ringkasan mempertahankan struktur dari teks referensi, dengan mempertimbangkan urutan kata yang muncul dalam teks. Relevansi Kata Kunci memastikan bahwa kata-kata kunci yang penting dari pertanyaan atau teks asli tetap ada dalam ringkasan. Skor BLEU dihitung dari persamaan berikut:

$$BLEU = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

di mana P_n adalah precision untuk n-gram ke-n, dan w_n adalah bobot untuk n-gram tersebut. Semakin tinggi nilai BLEU, semakin baik model dalam menghasilkan ringkasan yang sesuai dengan n-gram dari teks referensi. Kedua metrik ini memberikan gambaran yang komprehensif tentang seberapa baik kualitas ringkasan yang dihasilkan, baik dalam hal kesamaan leksikal (kata tunggal, pasangan kata) maupun kesamaan struktur dan relevansi kata kunci[3][4][5].

Result

Hasil dari penelitian memberikan hasil bahwa UnifiedQA-T5 menunjukkan perfoma yang paling baik dibandingkan dengan model lainnya seperti Msmarco-mt5, Cendol-mt5, T5-Large, mBart-large50, dan Pegasus dalam tugas question summarization pada bidang kesehatan dan berbahasa Indonesia. Hasil dari UnifiedQA-T5 tidak hanya lebih baik dalam hal menangkap kata kunci yang relevan (ROUGE-1: 0.2982) tetapi juga dalam konteks untuk menjaga struktur teks (ROUGE-L: 0.2492) dan menghasilkan ringkasan yang lebih sesuai dengan n-gram teks referensi (BLEU: 6.8375).

Berikut adalah hasil perbandingan model dari keenam model yang dipilih berdasarkan evaluasi menggunakan metrik ROUGE-1, ROUGE-2, ROUGE-L, dan BLEU Score.

Tabel 1. Performa Model

Metrics	Model					
	Msmarco-mt5	**Cendol-mt5**	**T5-Large**	**mBart-large50**	**Pegasus**	**Unifiedqa-T5**
ROUGE-1	0.2564	0.2397	0.2901	0.2134	0.0403	**0.2982**
ROUGE-2	0.0688	0.0659	0.0761	0.0673	0.0073	**0.0810**
ROUGE-L	0.2092	0.1941	0.2338	0.1798	0.0349	**0.2492**
BLEU	4.2615	3.4210	5.4499	3.0429	1.1245	**6.8375**

Berdasarkan hasil evaluasi yang ditampilkan pada tabel di atas, model Unifiedqa-T5 menunjukkan kinerja terbaik dibandingkan model-model lain, baik dari segi ROUGE maupun BLEU. Model ini mencapai skor ROUGE-1 tertinggi (0.2982), ROUGE-2 tertinggi (0.0810), dan ROUGE-L tertinggi (0.2492), serta BLEU tertinggi (6.8375). Hal ini menunjukkan bahwa Unifiedqa-T5 mampu menghasilkan ringkasan yang lebih mendekati kalimat referensi secara keseluruhan, baik dalam kemiripan leksikal maupun keselarasan semantik.

Di sisi lain, Pegasus mencatat kinerja paling rendah di semua metrik dengan selisih yang cukup signifikan, misalnya ROUGE-1 hanya 0.0403 dan BLEU 1.1245. Model-model lain seperti T5-Large, msmarco-mt5, cendol-mt5, dan mBart-large50 berada pada posisi menengah, dengan T5-Large cenderung lebih unggul di antaranya. Meskipun model-model tersebut memperlihatkan kemampuan yang cukup baik, khususnya T5-Large yang memiliki nilai ROUGE dan BLEU di atas rata-rata, tetap terlihat bahwa Unifiedqa-T5 memiliki tingkat keunggulan yang konsisten.

Setelah menemukan model yang paling baik, akan dilakukan *fine tuning* untuk mengoptimalkan model. Berikut adalah hasil performanya.

Tabel 2. Performa Model Terbaik Setelah *Fine Tuning*

Model	Metrics			
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BLEU**
Fine Tuned Unifiedqa-T5	**0.4992**	**0.28**	**0.4760**	0.1676

Performa model setelah dilakukan fine tuning meningkat tajam pada metric ROUGE-1, ROUGE-2, dan ROUGE-L. Sedangkan pada metric BLEU justru menurun tajam dari 6.8375 ke 0.1676.

Discussion

Berdasarkan Tabel 1, kinerja beberapa model dievaluasi menggunakan metrik ROUGE dan BLEU. Model Unifiedqa-T5 menunjukkan performa terbaik di hampir semua metrik, menandakan keunggulannya dibanding model lainnya. Pada metrik ROUGE-1, Unifiedqa-T5 memperoleh skor tertinggi sebesar 0.2982, melampaui T5-Large (0.2901) dan Msmarco-mt5 (0.2564). Untuk ROUGE-2, Unifiedqa-T5 juga unggul dengan skor 0.0810, mengalahkan T5-Large (0.0761) dan Msmarco-mt5 (0.0688). Pada ROUGE-L, Unifiedqa-T5 kembali menjadi

yang tertinggi dengan skor 0.2492, diikuti oleh T5-Large (0.2338). Sementara itu, pada metrik BLEU, Unifiedqa-T5 mencatat peningkatan signifikan dengan skor 6.8375, jauh di atas T5-Large (5.4499) dan Msmarco-mt5 (4.2615). Di sisi lain, model Pegasus menunjukkan performa terendah di seluruh metrik. Hasil ini menunjukkan bahwa Unifiedqa-T5 memiliki arsitektur dan strategi pelatihan yang lebih unggul, sehingga menghasilkan kinerja terbaik dalam mengevaluasi kesamaan kalimat menggunakan metrik ROUGE dan BLEU.

Setelah dilakukan fine-tuning pada model UnifiedQA-T5 di tabel 2, terdapat peningkatan signifikan dalam performa pada metrik ROUGE-1, ROUGE-2, dan ROUGE-L, yang menunjukkan bahwa *fine-tuning* membantu model menghasilkan ringkasan yang lebih relevan dan akurat. Nilai ROUGE-1 meningkat dari 0.2982 menjadi 0.4992, yang mengindikasikan peningkatan kemampuan model dalam memilih kata-kata yang relevan. ROUGE-2 dan ROUGE-L juga menunjukkan peningkatan, yang berarti *fine-tuning* membantu model lebih baik dalam menjaga struktur kalimat dan tumpang tindih pasangan kata. Namun, metrik BLEU mengalami penurunan drastis dari 6.8375 menjadi 0.1676, yang mungkin disebabkan oleh fokus *fine-tuning* pada konten yang relevan daripada kecocokan n-gram persis dengan referensi. Secara keseluruhan, *fine-tuning* meningkatkan kualitas ringkasan meskipun mengorbankan kecocokan n-gram yang ketat.

Berikut adalah contoh perbandingan hasil pertanyaan ringkasan dari semua model yang sudah diuji.

Tabel 3. Perbandingan Hasil Summary

	Referensi Teks	Hasil Summary Model
Contoh 1		
Msmarco-mt5	**Clean text:** adik saya kelas a5 sd pulang dari pondok pesantren jadi kutuan rambutnya ini bagaimana ya cara hilangin kutunya di apotek namanya apa obat pembasmi kutu rambut **Summary:** Obat pembasmi kutu rambut di apotek?	apakah ada obat pembasmi kutu di rambutmu jika kamu hilangnya kutunya a5 sd pulang pondok pesantren?
Cendol-mt5		kakak kelas a5 sd pulang pondok pesantren kutuan rambutnya ya hilangin kutuny apotek namanya obat pembasmi kutu rambutnya?
T5-Large		adik kelas sd pulang pondok pesantren kutuan rambutnya .?
mBart-large50		adik kelas a5 sd pulang pondok pesantren kutuan rambutnya ya hilangin kutunya apotek namanya obat pembasmi kutu rambut?
Pegasus		What is the difference between a 5 sd pulang pondok and a pesantren kutuan rambut in Indonesia, and what does it mean for you?
Unifiedqa-T5		adik kelas sd pulang pondok pesantren kutuan rambutnya?
Unifiedqa-T5 Fine tuned		Obat pembasmi kutu rambut di apotek
Contoh 2		
Msmarco-mt5	**Cleaned text:** bayi saya usia 9 bulan menangis tidak berhenti henti tangan kaki dingin wajah	apa yang menyebabkan bayi menangis berhenti henti tangan kaki dingin pucat badan dan wajah lemas di sekitar perut?
Cendol-mt5		bayi usia 9 menangis berhenti henti tangan kaki dingin wajah lemas pucat badan demam batuk pilek perutnya kembung menceret?
T5-Large		bayi usia 9 menangis berhenti tangan kaki dingin wajah lemas pucat badan demam batuk pilek perutnya kembung

	tampak lemas dan pucat badan tidak demam tidak batuk pilek juga namun perutnya kembung tidak menceret juga apakah yang saya sebutkan tadi ciri ciri bayi sakit perut dan cara mengatasinya	menceret sebutkan ciri baye .?
mBart-large50		bayi usia 9 menangis berhenti berhenti henti tangan kaki dingin wajah lemas pucat badan demam batuk pilek perutnya kembung menceret sebutkan ciri ciri bayi sakit perut mengatasinya apa yang harus kita lakukan jika bayi kita menangis?
Pegasus	**Summary:** Ciri-ciri bayi sakit perut dan cara mengatasinya	Is there a way to tell if a man who has been in a coma for 10 years is still alive?, asks the BBC's Mark Lowen.?
Unifiedqa-T5		bayi usia 9 menangis berhenti hente tangan kaki dingin wajah lemas pucat badan demam batuk pilek perutnya kembung menceret?
Unifiedqa-T5 Fine tuned		Ciri-ciri bayi sakit perut mengatasinya

Model MsMarco-mt5 menghasilkan ringkasan yang cukup baik meskipun tidak sepenuhnya menggambarkan inti dari pertanyaan, seperti pada Contoh 1, di mana ringkasan "apakah ada obat pembasmi kutu rambut jika kamu hilangnya kutunya" tidak mengarah pada informasi yang jelas mengenai obat yang dimaksud. Cendol-mt5 sedikit lebih jelas dengan menghasilkan teks yang lebih sesuai dengan pertanyaan yang diajukan, meskipun masih ada elemen yang kurang presisi, seperti pada "adik kelas a5 sd pulang pondok pesantren kutu rambutnya", yang seharusnya menjelaskan lebih rinci tentang kutu rambut dan obat yang dicari.

T5-Large memberikan hasil yang lebih baik dengan ringkasan yang lebih ringkas dan relevan, namun beberapa kata tetap muncul tidak tepat, seperti pada "Clean text: adik saya a5 sd pulang dari pondok pesantren" yang terlalu umum dan kurang mengarah pada pencarian solusi medis. Sedangkan mBart-large50 menghasilkan ringkasan yang lebih jelas, namun masih ada beberapa kalimat yang tidak terhubung langsung dengan inti pertanyaan, seperti "Ciri-ciri bayi sakit perut dan cara mengatasinya", yang terlalu umum dan kurang relevan dengan konteks aslinya. Pegasus memiliki hasil yang sedikit lebih relevan dengan ringkasan seperti "Summary: Obat pembasmi kutu rambut di apotek?", namun masih ada beberapa frasa yang terasa tidak langsung menjawab inti dari pertanyaan.

UnifiedQA-T5 menghasilkan ringkasan yang cukup akurat, meskipun masih terdapat kekurangan dalam keterbacaan teks, seperti pada "adik kelas a5 sd pulang pondok pesantren kutu rambutnya". Hasil yang di *fine-tune* menunjukkan peningkatan kualitas yang signifikan, dengan ringkasan yang lebih sesuai dan lebih relevan dengan pertanyaan yang diajukan, seperti "Ciri-ciri bayi sakit perut mengatasinya". Ini menunjukkan bahwa *fine-tuning* pada UnifiedQA-T5 memberi dampak positif pada kualitas ringkasan, memungkinkan model menghasilkan teks yang lebih tepat dan informatif.

Setelah melakukan evaluasi terhadap hasil ringkasan yang dihasilkan oleh model, selanjutnya mengeksplorasi pengaruh penggunaan *stopwords* terhadap kinerja model. Lalu, membandingkan hasil ringkasan yang dihasilkan oleh model baik dengan menggunakan *stopwords* maupun tanpa *stopwords* untuk melihat sejauh mana pengaruhnya terhadap akurasi dan kualitas ringkasan yang dihasilkan.

Tabel 4. Perbandingan Performa dengan dan Tanpa Stopwords

Model		ROGUE-1	ROGUE-2	ROGUE-L	BLEU
Msmarco-mt5	Stopword	0.2558	0.0680	0.2070	4.1956
	Tanpa Stopword	**0.2564**	**0.0688**	**0.2092**	**4.2615**
Cendol-mt5	Stopword	0.2375	0.0658	0.1920	**3.4285**
	Tanpa Stopword	**0.2397**	**0.0659**	**0.1941**	3.2410
T5-Large	Stopword	0.2899	0.0744	**0.2340**	5.4197
	Tanpa Stopword	**0.2901**	**0.0761**	0.2338	**5.4499**
mBart-large50	Stopword	0.2121	0.0658	0.1775	2.9769
	Tanpa Stopword	**0.2134**	**0.0673**	**0.1793**	**3.0429**
pegasus	Stopword	**0.0414**	**0.0074**	**0.0365**	**1.1285**
	Tanpa Stopword	0.0403	0.0073	0.0349	1.1245
Unifiedqa-T5	Stopword	0.2706	**0.1119**	0.2349	5.4137
	Tanpa Stopword	**0.2982**	0.0810	**0.2496**	**6.8375**

Berdasarkan Tabel 4, perbandingan hasil model dengan dan tanpa penggunaan stopwords menunjukkan dampak yang signifikan terhadap kinerja model dalam menghasilkan ringkasan. Secara umum, penggunaan stopwords cenderung menghasilkan ROUGE-1, ROUGE-2, dan ROUGE-L yang lebih rendah pada sebagian besar model, dengan BLEU score yang juga menunjukkan tren serupa. Model seperti MsMarco-mt5, Cendol-mt5, dan mBart-large50 menunjukkan sedikit penurunan dalam nilai ROUGE dan BLEU ketika menggunakan stopwords dibandingkan tanpa *stopwords*. Namun, UnifiedQA-T5 menunjukkan hasil yang lebih baik dengan stopwords, menghasilkan nilai ROUGE-1 sebesar 0.2982 dan BLEU 6.8375, yang lebih tinggi dibandingkan dengan tanpa *stopwords*. Ini menunjukkan bahwa penggunaan *stopwords* dapat meningkatkan kinerja model tertentu, seperti UnifiedQA-T5, sementara pada model lainnya, penghapusan *stopwords* justru memberikan hasil yang lebih baik dalam hal kecocokan n-gram dan kesesuaian dengan referensi.

Tabel 5. Perbandingan Contoh Hasil Summary dengan dan Tanpa Stopwords

Clean Text: mau bertanya anak saya usia 3 bulan asi ekslusif lidahnya berwarna putih saya sulit membedakan putihnya dikarenakan warna asi atau ada jamur lidah sejauh ini anak saya tidak begitu rewel dan masih mau menyusui meskipun tidak sesering biasanya saya sebagai orang tua bingung bagaimana ya cara untuk menghilangkan bercak putih tersebut
Summary: Cara menghilangkan bercak putih di bibir bayi

Model	Hasil Pertanyaan	
	Stopword	**Tanpa Stopword**
Msmarco-mt5	kenapa lidahmu berwarna putih saatpada usia berapa anak rewel menyusui menyusui setelah usia 3 bulan sayabisa membedakan warna lidahnya tidak bisa membedakan warnanyadengan asi ekslusif seperti jamur?	

	dengan asi ekslusif?	
Cendol-mt5	Cara untuk menghilangkan bercak putih ini adalah sebagai berikut: 1. Menghilangkan Bercak Putih Jika Anda sudah berusia 3 bulan, jangan begitu sering menyusui. Jangan takut jika Anda tidak merasa sedih. 2. Menghapus Keputihan Wanita Coba gunakan obat-obatan yang dibutuhkan oleh orang tua Anda. Biarkan anak Anda mengalami cedera. 3. Mengatasi Kemerahan Kulit Berwarna Cerah Periksa apakah Anda memiliki tubuh yang lebih ringan. Pahami bagaimana Anda bisa menggunakan bahan-bahan yang Anda butuhkan. Pertimbangkan untuk mengurangi warna-warna seperti?	Anak usia 3 asi ekslusif lidah anak rewel menyusui sesering orang tua bingung ya menghilangkan bercak putih?
T5-Large	mau bertanya anak saya usia 3 bulan asi ekslusif lidahnya berwarna putih tsb .?	anak rewel menyusui sesering orang tua bingung ya menghilangkan bercak putih sulit .?
mBart-large50	mau bertanya anak saya usia 3 bulan asi ekslusif lidahnya berwarna putih saya sulit membedakan putihnya dikarenakan warna asi atau ada jamur di lidah saya tidak begitu rewel dan masih mau menyusui meskipun tidak sesering biasanya saya sebagai orang tua bingung bagaimana ya cara ...?	anak usia 3 asi ekslusif lidahnya berwarna putih sulit membedakan putihnya warna asi jamur lidah anak rewel menyusui sesering orang tua bingung ya menghilangkan bercak putih di mulutnya, apa yang harus saya lakukan?
PEGASUS	Do you think the Indonesian government is doing enough to protect the rights of Orangutans and other wildlife species in Indonesia, especially those living on the island of Borneo?	Is it possible for the Indonesian government to put a lidah on the port of Medan, which has been a major issue in the country for years?
Unifiedqa-T5	warna asi atau ada jamur lidah sejauh ini anak saya tidak begitu rewel?	warna asi jamur lidah anak rewel menyusui sesering orang tua bingung?

Berdasarkan perbandingan hasil ringkasan yang dihasilkan dengan *stopwords* dan tanpa stopwords pada setiap model di tabel 5, terlihat bahwa penggunaan *stopwords* berpengaruh terhadap kualitas ringkasan yang dihasilkan. Pada MsMarco-mt5, ringkasan yang dihasilkan dengan *stopwords* cenderung mengandung kata-kata yang tidak relevan seperti "lidahmu", sedangkan tanpa *stopwords*, model menghasilkan ringkasan yang lebih langsung dan tepat sasaran. Begitu juga dengan Cendol-mt5, di mana penggunaan *stopwords* menghasilkan ringkasan yang lebih panjang dan kurang fokus, sementara tanpa *stopwords* menghasilkan ringkasan yang lebih sederhana dan tepat meskipun ada beberapa informasi yang hilang. Untuk

T5-Large, penggunaan *stopwords* memberikan ringkasan yang lebih terstruktur dan informatif, sedangkan tanpa *stopwords* menghasilkan ringkasan yang lebih ringkas namun kehilangan beberapa kata kunci penting.

MBart-large50 menunjukkan hasil yang serupa, di mana dengan *stopwords* ringkasan menjadi lebih jelas namun panjang, sementara tanpa *stopwords* ringkasan menjadi lebih singkat dan sedikit kehilangan informasi. PEGASUS menunjukkan bahwa penggunaan *stopwords* mengarah pada ringkasan yang lebih luas, tetapi agak menyimpang dari konteks, sementara tanpa *stopwords* ringkasan menjadi lebih fokus pada inti pertanyaan. Terakhir, UnifiedQA-T5 menghasilkan ringkasan yang lebih rinci dengan *stopwords*, namun beberapa informasi tambahan menjadi tidak perlu. Tanpa *stopwords*, ringkasan menjadi lebih efisien dan langsung pada inti masalah. Secara keseluruhan, meskipun setiap model menunjukkan perbedaan, tanpa *stopwords* cenderung memberikan ringkasan yang lebih tepat dan relevan, kecuali pada beberapa model tertentu yang lebih efektif dengan *stopwords* untuk menjaga struktur dan konteks.

Untuk menilai hasil secara kuantitatif, metode akan dibandingkan melalui penilaian manusia (*human judgement*). Labeling manual dilakukan pada 1000 data secara acak, dengan mengkategorikan setiap ringkasan yang dihasilkan ke dalam salah satu kategori berikut: 2 (*Perfect Summary*), 1 (*Acceptable Summary*), dan 0 (*Incorrect Summary*). Rata-rata jumlah untuk setiap kategori dihasilkan pada Tabel 4.

Tabel 6. Proporsi Human Judgement

Mutu	Proporsi
0 (Incorrect Summary)	14. 8
1 (Acceptable Summary)	33.8
2 (Perfect Summary)	51.4

Hasil evaluasi menunjukkan bahwa model *fine tuned* UnifiedQA-T5 berhasil menghasilkan 51.4% ringkasan sempurna, yang menandakan bahwa model ini efektif dalam memahami dan merangkum pertanyaan kesehatan dengan baik. Selain itu, 33.8% dari ringkasan berada dalam kategori Ringkasan Diterima (*Acceptable Summary*), yang berarti sekitar sepertiga dari hasil ringkasan masih dapat diterima meskipun mungkin ada beberapa kekurangan atau ketidaktepatan dalam informasi yang disajikan. Namun, meskipun sebagian besar ringkasan dapat diterima atau sempurna, masih ada sekitar 14.8% ringkasan yang salah, yang menunjukkan ada ruang untuk perbaikan dalam akurasi model, terutama dalam menangani kompleksitas dan keakuratan informasi yang lebih rinci.

Tabel 7. Contoh Hasil Summary Pertanyaan berdasarkan Mutu

No	ROGUE-1 Category	Hasil Pertanyaan	ROGUE-1 Score
357	Category 0	Bagaimana cara menanganinya gimana?	0.000000
984		Cara mengatasi masuk angin sembuh	0.307692
577		Bagaimana cara mengatasi biduran setelah mandi?	0.333333
856	Category 1	Cara mengatasi siku tangan sebelah kanan	0.461538
361		Bagaimana cara mengganjal sesak napas dan faringitis?	0.461538

932		Benjolan di gusi rahang dan benjolan di lidah	0.400000
571	Category 2	Cara terhindar dari cedera berolahraga	0.909091
973		Apa saja vitamin patah tulang apotik?	0.666667
847		Benjolan dibelakang telinga bayi	0.615385

Berdasarkan Tabel 7, dapat dilihat perbandingan ROUGE-1 *score* untuk berbagai kategori hasil ringkasan. Pertanyaan yang termasuk dalam Kategori 0 (Salah), seperti "Bagaimana cara menanggani gimana?", memiliki ROUGE-1 *score* yang sangat rendah (0.000000), menunjukkan bahwa model tidak berhasil menghasilkan ringkasan yang relevan atau sesuai dengan teks referensi. Sementara itu, pertanyaan dalam Kategori 1 (Diterima), seperti "Bagaimana cara mengatasi siku tangan sebelah kanan?" dan "Bagaimana cara mengatasi sesak napas dan faringitis?", memiliki ROUGE-1 *score* sekitar 0.461538, yang mengindikasikan bahwa model menghasilkan ringkasan yang cukup baik meskipun masih terdapat beberapa ketidaktepatan dalam kesesuaian kata-kata atau informasi. Terakhir, pertanyaan dalam Kategori 2 (Sempurna), seperti "Apa saja vitamin patah tulang apotik?" dan "Benjolan di belakang telinga bayi", memiliki ROUGE-1 *score* yang lebih tinggi, yaitu antara 0.615385 hingga 0.666667, menunjukkan bahwa model sangat efektif dalam menghasilkan ringkasan yang sangat relevan dan sesuai dengan referensi. Secara keseluruhan, hasil ini menunjukkan bahwa model menunjukkan kinerja terbaik pada kategori yang lebih relevan (Kategori 2) dan kesulitan dalam menghasilkan ringkasan yang tepat pada kategori yang lebih sulit atau ambigu (Kategori 0).

Conclusion

Penelitian ini telah berhasil mengembangkan model *summarization* berbasis transfer learning untuk merangkum pertanyaan panjang dalam konteks kesehatan konsumen di platform *telemedicine* berbahasa Indonesia. Evaluasi menunjukkan bahwa model UnifiedQA-T5, setelah melalui proses *fine-tuning*, memberikan performa terbaik dalam menghasilkan ringkasan yang relevan, terstruktur, dan akurat. Model ini menunjukkan peningkatan signifikan pada skor ROUGE-1 (0.2982 ke 0.4992), ROUGE-2 (0.0810 ke 0.2800), dan ROUGE-L (0.2492 ke 0.4760), meskipun BLEU menurun dari 6.8375 ke 0.1676, yang mengindikasikan pergeseran fokus ke relevansi konten daripada kecocokan n-gram.

Model ini juga menunjukkan kemampuan adaptasi yang baik terhadap karakteristik bahasa Indonesia, seperti fleksibilitas urutan kata dan variasi kosakata, yang sering menjadi tantangan dalam tugas pemrosesan bahasa alami. Selain itu, penerapan *fine-tuning* memungkinkan model untuk menangkap makna inti dari pertanyaan secara lebih efektif, meskipun terdapat penurunan pada beberapa aspek evaluasi kuantitatif. Model ini dapat menjadi solusi untuk memahami pertanyaan panjang yang kompleks, yang sering kali dipengaruhi oleh struktur bahasa, variasi kosakata, serta keterbatasan dataset. Meskipun hasilnya sangat menjanjikan, model ini masih memiliki kekurangan, di mana 14,8% ringkasan berada dalam kategori tidak sesuai, 33,8% dalam kategori cukup diterima, dan hanya 51,4% yang masuk kategori sempurna. Hal ini menunjukkan perlunya pengembangan lebih lanjut untuk meningkatkan akurasi dan konsistensi dalam menghasilkan ringkasan yang berkualitas.

Future Work

Untuk penelitian selanjutnya, disarankan beberapa langkah untuk meningkatkan performa model:

1. Melakukan eksplorasi lebih lanjut terhadap *ensemble methods* dengan menggabungkan beberapa model dan pendekatan yang berbeda untuk meningkatkan akurasi dan generalisasi.
2. Melakukan *hyperparameter-tuning* juga dapat dilakukan untuk mengoptimalkan kinerja model, dengan menyesuaikan berbagai parameter penting seperti *learning rate*, jumlah lapisan, dan ukuran batch, untuk menemukan kombinasi terbaik yang dapat meningkatkan kualitas ringkasan.
3. Selain metrik kuantitatif seperti ROUGE dan BLEU, evaluasi kualitatif oleh ahli bahasa atau profesional medis dapat memberikan insight tambahan mengenai kualitas ringkasan yang dihasilkan.

References

Abacha, A. ben, & Demner-Fushman, D. (2019). *A Question-Entailment Approach to Question Answering*. https://doi.org/10.1186/s12859-019-3119-4

Abacha, A. ben, Mrabet, Y., Zhang, Y., Shivade, C., Langlotz, C., & Demner-Fushman, D. (2021). *Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain*. https://bionlp.nlm.nih.gov/

Akhir, T. (n.d.). *Peringkas Teks Otomatis Bahasa Indonesia secara Abstraktif Menggunakan Metode Long Short-Term Memory*.

Balumuri, S., Bachina, S., & Kamath, S. (2021). *SB_NITK at MEDIQA 2021: Leveraging Transfer Learning for Question Summarization in Medical Domain*.

Bahari, A., & Dewi, K. E. (2024). *PERINGKASAN TEKS OTOMATIS ABSTRAKTIF MENGGUNAKAN TRANSFORMER PADA TEKS BAHASA INDONESIA*. *13*(1).

Ben Abacha, A., & Demner-Fushman, D. (2019). On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2228–2234). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1215

Dash, Y., Kumar, A., Chauhan, S. S., Singh, A. V., Ray, A., & Abraham, A. (2024). Advances in medical text summarization: Comparative performance analysis of PEGASUS and T5. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCCNT61001.2024.10724845

Ghafouri, A., Barati, I., Elahimanesh, M. H., & Hasanpour, H. (2023). A Question Summarization Method based-on Deep Learning in Persian Language. *2023 28th International Computer Conference, Computer Society of Iran, CSICC 2023*. https://doi.org/10.1109/CSICC58665.2023.10105324

Ghosh, A., Acharya, A., Jain, R., Saha, S., Chadha, A., & Sinha, S. (2024). CLIPSyntel: CLIP and LLM synergy for multimodal question summarization in healthcare. *AAAI, 38*(20), 22031–22039. https://doi.org/10.1609/aaai.v38i20.30206

Karo, I. M. K., Perdana, A., & Dewi, S. (2024). Automatic Text Review Summarization of Digital Library System Application using TextRank Algorithm and TF-IDF. *2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 570–575. https://doi.org/10.1109/ICSINTESA62455.2024.10747952

Lee, J., Dang, H., Uzuner, O., & Henry, S. (2021). MNLP at MEDIQA 2021: Fine-tuning PEGASUS for consumer health question summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 320–327). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.37

Lee, L.-H., Chen, P.-H., Zeng, Y.-X., Lee, P.-L., & Shyu, K.-K. (2021). NCUEE-NLP at MEDIQA 2021: Health question summarization using PEGASUS transformers. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 268–272). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.30

Lubis, A. R., Safitri, H. R., Irvan, Lubis, M., Hamzah, M. L., Al-Khowarizmi, A. K., & Nugroho, O. (2023). Enhancing Text Summarization with a T5 Model and Bayesian Optimization. *Revue d'Intelligence Artificielle*, *37*(5), 1213–1219. https://doi.org/10.18280/ria.370513

Mrini, K., Dernoncourt, F., Chang, W., Farcas, E., & Nakashole, N. (2021). Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations* (pp. 58–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nlpmc-1.8

Nasari, M., Maulina, A., & Girsang, A. S. (2023). Abstractive Indonesian News Summarization Using BERT2GPT. *Proceedings - 2023 IEEE 7th International Conference on Information*

Technology, Information Systems and Electrical Engineering, ICITISEE 2023, 369–375. https://doi.org/10.1109/ICITISEE58992.2023.10405359

Primakara, U. (n.d.). *IMPLEMENTASI PERINGKAS DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN METODE TEXT TO TEXT TRANSFER TRANSFORMER (T5) I Nyoman Purnama 1) , Ni Nengah Widya Utami 2) Program Studi Sistem Informasi 1) , Sistem Informasi Akutansi 2)*.

Ridho Lubis, A., Ramdani Safitri, H., & Lubis, M. (n.d.). IMPROVING TEXT SUMMARIZATION QUALITY BY COMBINING T5-BASED MODELS AND CONVOLUTIONALSEQ2SEQ MODELS. In *Journal of Applied Engineering and Technological Science* (Vol. 5, Issue 1).

Sänger, M., Weber, L., & Leser, U. (2021). WBI at MEDIQA 2021: Summarizing consumer health questions with generative transformers. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 86–95). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.9

Savery, M., Abacha, A. B., Gayen, S., & Demner-Fushman, D. (2020). Question-driven summarization of answers to consumer health questions. *Scientific Data, 7*(1), 322. https://doi.org/10.1038/s41597-020-00667-z

Wang, X., Wang, J., Xu, B., Lin, H., Zhang, B., & Yang, Z. (2022). Exploiting intersentence information for better question-driven abstractive summarization: Algorithm development and validation. *JMIR Medical Informatics, 10*(8), e38052. https://doi.org/10.2196/38052

Wei, S., Lu, W., Peng, X., Wang, S., Wang, Y.-F., & Zhang, W. (2023). Medical question summarization with entity-driven contrastive learning. *arXiv*. https://doi.org/10.48550/arXiv.2304.07437

Widodo, W., Nugraheni, M., & Sari, I. P. (2021). A comparative review of extractive text summarization in Indonesian language. *IOP Conference Series: Materials Science and Engineering*, *1098*(3), 032041. https://doi.org/10.1088/1757-899x/1098/3/032041

Wijayanti, R., Khodra, M. L., & Widyantoro, D. H. (2021). Indonesian Abstractive Summarization using Pre-Trained Model. *3rd 2021 East Indonesia Conference on*

Computer and Information Technology, EIConCIT 2021, 79–84. https://doi.org/10.1109/EIConCIT50028.2021.9431880

Wilson, A., & Zahra, A. (2022). IMPROVING TEXT FEATURES IN TEXT SUMMARIZATION USING HARRIS HAWKS OPTIMIZATION. *ICIC Express Letters*, *16*(9), 923–932. https://doi.org/10.24507/icicel.16.09.923

Yadav, S., Gupta, D., Abacha, A. ben, & Demner-Fushman, D. (2021). *Reinforcement Learning for Abstractive Question Summarization with Question-aware Semantic Rewards*. http://arxiv.org/abs/2107.00176

Yadav, S., Gupta, D., Abacha, A. B., & Demner-Fushman, D. (2021). Question-aware transformer models for consumer health question summarization. *arXiv*. https://doi.org/10.48550/arXiv.2106.00219

Yadav, S., Sarrouti, M., & Gupta, D. (2021). NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 291–301). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.34

Yadav, S., Gupta, D., Abacha, A. B., & Demner-Fushman, D. (2022). Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics, 128*, 104040. https://doi.org/10.1016/j.jbi.2022.104040

Yadav, S., & Caragea, C. (2024). Towards summarizing healthcare questions in low-resource setting.

Zhang, M., Dou, S., Wang, Z., & Wu, Y. (2024). Focus-driven contrastive learning for medical question summarization.

Zhang, L., & Liu, J. (2022). Intent-aware prompt learning for medical question summarization. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 672–679). IEEE. https://doi.org/10.1109/BIBM55620.2022.9995317

Zekaoui, N. E., Yousfi, S., Mikram, M., & Rhanoui, M. (2023). Enhancing large language models' utility for medical question-answering: A patient health question summarization approach. In *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1–8). IEEE. https://doi.org/10.1109/SITA60746.2023.10373720