Pengembangan Model BERT dengan Metode *Abstraction Summarization* untuk Ringkasan Otomatis Tafsir Ayat Al-Qur'an

Proposal Skripsi

diajukan sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer

oleh
Reiki Aziz Yoga Utama
4611422055

TEKNIK INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI SEMARANG
SEMARANG
2024

PERSETUJUAN PEMBIMBING

Skripsi berjudul "Pengembangan Model BERT dengan Metode *Abstraction Summarization* untuk Ringkasan Otomatis Tafsir Ayat Al-Qur'an" yang disusun oleh :

nama	: Reiki Aziz Yoga Utama
NIM	: 4611422055
Prodi/Fakultas	: Teknik Informatika/Matematika dan Ilmu Pengetahuan Alam

Semarang, 06 September 2024
Belum ditentukan

Belum ditentukan & NIP

BAB I

PENDAHULUAN

1.1 Latar Belakang

Al-Qur'an merupakan kitab suci umat Islam yang menjadi sumber hukum islam yang pertama dan utama (Aji Fitra Jaya Institut Perguruan Tinggi Ilmu Al Qur, n.d.). Dalam praktiknya, pemahaman terhadap Al-Qur'an seringkali membutuhkan penjelasan atau tafsir yang mendalam agar pesan yang terkandung dapat dimengerti oleh berbagai kalangan. Tafsir Al-Qur'an adalah karya ilmiah yang menjelaskan makna ayat-ayat Al-Qur'an secara mendetail, baik dalam bahasa Arab maupun bahasa lainnya, termasuk bahasa Indonesia. Namun, panjangnya tafsir sering kali menjadi tantangan, terutama bagi mereka yang membutuhkan pemahaman cepat namun tetap akurat.

Seiring dengan perkembangan teknologi, berbagai pendekatan berbasis kecerdasan buatan (Artificial Intelligence) mulai diterapkan untuk mempermudah akses terhadap informasi, termasuk dalam bidang teks keagamaan. Salah satu teknologi yang berperan penting dalam pengolahan teks adalah **Natural Language Processing (NLP)**. Dalam konteks ini, **summarization** atau teknik peringkasan otomatis merupakan salah satu pendekatan NLP yang dapat membantu meringkas tafsir tanpa kehilangan esensi dari teks aslinya. Summarization menawarkan kemudahan bagi pengguna untuk memahami inti dari tafsir yang panjang dalam waktu yang lebih singkat.

Pada umumnya, terdapat dua metode utama dalam summarization, yaitu **extractive** dan **abstractive summarization**. **Extractive summarization** mengekstraksi kalimat-kalimat penting dari teks asli, sedangkan **abstractive summarization** mencoba meringkas informasi dengan menyusun kalimat baru yang lebih singkat dan koheren. Abstractive summarization memiliki keunggulan dalam menciptakan ringkasan yang lebih alami dan tidak sekadar mengutip langsung teks asli, namun memiliki tantangan lebih besar dalam memahami konteks yang kompleks (Alami Merrouni et al., 2023). Dalam konteks teks bilingual seperti tafsir Al-Qur'an dalam bahasa Arab dan Indonesia, abstractive summarization memberikan potensi yang lebih besar untuk menghasilkan ringkasan yang informatif dan tidak bergantung pada struktur kalimat asli, sehingga cocok untuk menangani kompleksitas dan perbedaan bahasa tersebut.

Salah satu model yang dapat mendukung penerapan **abstractive summarization** adalah **BERT** (Bidirectional Encoder Representations from Transformers). BERT telah terbukti sangat efektif dalam berbagai tugas NLP, termasuk **abstractive summarization**, karena kemampuannya untuk memahami konteks dua arah secara mendalam dalam teks. Ini memungkinkan BERT untuk menghasilkan ringkasan yang lebih akurat dan koheren, terutama untuk teks panjang dan kompleks. Namun, penggunaan model BERT dalam konteks tafsir Al-Qur'an yang bilingual (Arab-Indonesia) masih belum banyak dieksplorasi. Padahal, potensinya sangat besar untuk meningkatkan aksesibilitas dan pemahaman terhadap tafsir dengan menghasilkan ringkasan otomatis yang lebih informatif dan mudah dipahami.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk mengoptimalkan penggunaan metode **abstractive summarization** berbasis model transformer **BERT** dalam meringkas tafsir Al-Qur'an dalam dua bahasa, yaitu Arab dan Indonesia. Dengan memanfaatkan teknologi ini, diharapkan hasil ringkasan dapat tetap mempertahankan makna dan esensi tafsir yang kompleks, namun disajikan dalam bentuk yang lebih ringkas dan mudah dipahami oleh pengguna. Penelitian ini juga diharapkan dapat mengisi kesenjangan penelitian sebelumnya yang belum sepenuhnya memanfaatkan potensi summarization berbasis transformer untuk teks religius bilingual, khususnya dalam konteks tafsir Al-Qur'an.

1.2 Batasan Masalah

Penelitian ini memiliki beberapa batasan untuk memastikan fokus yang jelas dan pencapaian tujuan yang diinginkan. Pertama, penelitian ini difokuskan hanya pada teks tafsir Al-Qur'an yang berbahasa Arab serta terjemahannya dalam bahasa Indonesia, sehingga tidak mencakup teks keagamaan lain atau teks tafsir dalam bahasa selain Arab dan Indonesia. Kedua, penelitian ini menggunakan metode **abstractive summarization** berbasis transformer **BERT**, tanpa menerapkan metode lain di luar pendekatan ini. Model yang digunakan terbatas pada BERT (Bidirectional Encoder Representations from Transformers) yang disesuaikan untuk menangani teks bilingual (Arab-Indonesia), dan tidak mengeksplorasi model lain seperti GPT, T5, atau model transformer lainnya.

Selain itu, evaluasi kinerja dilakukan menggunakan metrik standar seperti ROUGE, tanpa menggunakan metrik evaluasi lain yang lebih kompleks atau spesifik terhadap domain tertentu. Dataset yang digunakan juga dibatasi pada kumpulan tafsir yang telah tersedia secara publik dan terjemahannya yang resmi, sehingga tidak dilakukan pengumpulan atau pembuatan dataset baru dari sumber yang belum terverifikasi. Akhirnya, penelitian ini hanya mempertimbangkan konteks teks pada level kalimat dan paragraf dalam tafsir, tanpa memasukkan konteks historis atau sosial dari tafsir Al-Qur'an secara keseluruhan.

1.3 Rumusan Masalah

Berdasarkan latar belakang dan batasan yang telah dijelaskan, rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana menerapkan metode **Abstractive summarization** berbasis transformer BERT untuk menghasilkan ringkasan otomatis dari teks bilingual (Arab-Indonesia) pada tafsir Al-Qur'an?
2. Sejauh mana abstractive summarization mampu menangkap esensi dari tafsir Al-Qur'an dalam dua bahasa?
3. Bagaimana performa model abstractive summarization dalam meringkas tafsir Al-Qur'an bila dievaluasi menggunakan metrik seperti ROUGE?

1.4 Tujuan Penelitian

Tujuan penelitian ini dapat dirinci sebagai berikut:

1. Menerapkan metode *abstraction summarization* berbasis transformer BERT untuk menghasilkan ringkasan otomatis dari teks bilingual (Arab-Indonesia) pada tafsir Al-Qur'an.
2. Sejauh mana pendekatan **abstractive summarization** mampu menangkap esensi dari tafsir Al-Qur'an dalam dua bahasa?.
3. Bagaimana performa model **abstractive summarization** dalam meringkas tafsir Al-Qur'an bila dievaluasi menggunakan metrik seperti ROUGE?

1.5 Manfaat Penelitian

1. **Kontribusi Teoritis**:
    - Mengembangkan dan memperkaya metode **abstractive summarization** berbasis transformer BERT untuk teks bilingual (Arab-Indonesia), khususnya dalam konteks tafsir Al-Qur'an.
    - Menyediakan wawasan baru dalam literatur pemrosesan bahasa alami (NLP) dan summarization dengan penerapan teknologi canggih pada teks keagamaan.
2. **Manfaat Praktis**:
    - Mempermudah proses meringkas teks tafsir Al-Qur'an, membuat informasi lebih aksesibel dan mudah dipahami oleh pembaca bilingual.
    - Menyediakan model summarization yang dapat digunakan oleh lembaga penelitian, pengembang aplikasi keagamaan, dan institusi pendidikan untuk menyederhanakan informasi dari tafsir dan meningkatkan kualitas bahan ajar serta sumber daya belajar.
3. **Aplikasi Potensial**:
    - Model yang dikembangkan dapat diterapkan dalam aplikasi edukasi dan penelitian untuk memfasilitasi pemahaman yang lebih baik terhadap tafsir Al-Qur'an.
    - Meningkatkan efisiensi dalam pengembangan sumber daya belajar dan materi ajar yang berkaitan dengan teks keagamaan.

1.6 Kebaruan Penelititan

Kebaruan penelitian ini terletak pada penerapan **abstractive summarization** berbasis transformer **BERT** dalam meringkas teks bilingual (Arab-Indonesia) pada tafsir Al-Qur'an. Dari segi konsep, penelitian ini memperkenalkan penggunaan pendekatan **abstractive** untuk meringkas teks keagamaan, khususnya tafsir Al-Qur'an, yang belum banyak dieksplorasi. Pendekatan ini menawarkan cara baru dalam menangani teks religius yang kompleks dan bilingual, yang bertujuan menghasilkan ringkasan yang lebih alami dan koheren. Dari segi metode, penelitian ini mengadopsi transformer **BERT** yang disesuaikan untuk menangani teks dalam dua bahasa secara bersamaan. Hal ini memungkinkan BERT untuk memahami

konteks bahasa Arab dan Indonesia secara efektif, sehingga dapat menghasilkan ringkasan yang lebih akurat dan informatif. Inovasi ini memperluas cakupan aplikasi model transformer dalam pemrosesan teks religius bilingual, memberikan metode baru dalam menghasilkan ringkasan otomatis yang efektif dan relevan.

BAB II

Kajian Pustaka

2.1 Tinjauan Pustaka

Dalam penelitian mengenai **abstractive summarization** berbasis model **BERT** untuk teks bilingual (Arab-Indonesia) dalam konteks **tafsir Al-Qur'an**, berbagai kajian sebelumnya dapat dijadikan landasan penting. Secara umum, summarization terbagi dalam dua metode utama: **extractive** dan **abstractive**. Pada extractive summarization, kalimat-kalimat penting dipilih dari teks asli tanpa perubahan struktur, sedangkan abstractive summarization menghasilkan kalimat baru berdasarkan pemahaman semantik terhadap teks.

Penelitian oleh (Nallapati et al., n.d.) menggarisbawahi efektivitas metode extractive dalam menghasilkan ringkasan yang cepat, tetapi dengan keterbatasan dalam fleksibilitas dan akurasi semantik. Sebaliknya, penelitian oleh (Liu & Lapata, n.d.) menunjukkan bahwa abstractive summarization berbasis **BERT** memiliki kemampuan lebih baik dalam menghasilkan ringkasan yang lebih menyerupai interpretasi manusia, karena mampu memahami konteks dan makna dari teks yang lebih luas.

Selain itu, penelitian oleh (Alselwi & Taşcı, 2024) mengembangkan teknik summarization berbasis graf untuk teks Arab dengan menggunakan **word embedding** dan algoritma **PageRank**, yang dikenal sebagai **GEATS (Graph-based Extractive Arabic Text Summarization)**. Teknik ini memanfaatkan pendekatan berbasis graf untuk menangani hubungan morfologis yang kompleks dalam bahasa Arab. Hasil dari penelitian ini menunjukkan bahwa pendekatan GEATS memberikan hasil ringkasan yang lebih baik dibandingkan metode lainnya, dengan peningkatan lebih dari 7,5% pada nilai **F-measure**.

Penelitian-penelitian tersebut menjadi fondasi penting untuk pengembangan model summarization berbasis **BERT** dalam konteks teks bilingual yang lebih spesifik seperti tafsir Al-Qur'an, di mana makna dan esensi dari teks keagamaan harus tetap terjaga dalam proses peringkasan.

2.2 Landasn Teoritik

Landasan teoretik penelitian ini bertumpu pada teori summarization, **Natural Language Processing (NLP)**, dan model **transformer** seperti BERT. Summarization adalah proses menyajikan informasi dalam bentuk yang lebih singkat namun tetap menyampaikan esensi teks. Dalam NLP, **abstractive**

summarization adalah metode yang menghasilkan ringkasan baru berdasarkan pemahaman semantik model terhadap teks. **BERT** (Bidirectional Encoder Representations from Transformers) telah terbukti efektif dalam memahami konteks teks dengan mempelajari hubungan antara kata-kata dalam dua arah (dua arah, atau bidirectional), seperti yang diuraikan oleh (Devlin et al., n.d.). BERT memungkinkan penggunaan konteks yang lebih kaya dalam proses summarization, yang penting dalam teks kompleks seperti tafsir Al-Qur'an yang mengandung makna mendalam.

Teks bilingual dalam penelitian ini, yakni Arab dan Indonesia, menambah lapisan kompleksitas karena BERT harus memahami dua bahasa yang berbeda secara struktur dan gramatikal. Penelitian ini menggunakan BERT yang telah ditransfer ke domain teks keagamaan dan teks bilingual, dengan abstractive summarization untuk meringkas tafsir Al-Qur'an.

2.3 Kerangka Berpikir

Kerangka berpikir penelitian ini mengacu pada gagasan bahwa teks tafsir Al-Qur'an yang bilingual (Arab-Indonesia) memiliki struktur dan makna yang kompleks, sehingga memerlukan metode summarization yang mampu menangkap esensi dalam dua bahasa. Model **abstractive summarization** berbasis BERT dipilih karena mampu menangani konteks teks dua arah dan menggabungkan makna dari berbagai segmen teks. Model ini akan dilatih pada teks bilingual untuk memahami hubungan semantik antarbahasa, dengan output berupa ringkasan yang informatif dan akurat. Pendekatan ini diharapkan mampu menjawab tantangan dalam menyajikan ringkasan yang ringkas namun tidak kehilangan esensi tafsir.

2.4 Hipotesis Teoritis

Penelitian ini mengajukan hipotesis bahwa **abstractive summarization** berbasis **BERT** dapat menghasilkan ringkasan yang lebih koheren, akurat, dan informatif dibandingkan dengan metode summarization lainnya untuk teks tafsir Al-Qur'an dalam dua bahasa. Model yang dihasilkan diharapkan dapat menangkap esensi dari tafsir Al-Qur'an, dan performanya akan dievaluasi menggunakan metrik standar seperti ROUGE.

BAB III

Metode Penelitian

3.1 Pendekatan, Jenis, dan Prosedur penelitian

Penelitian ini menggunakan pendekatan **kuantitatif** berbasis **eksperimen** untuk mengukur efektivitas model **hybrid summarization** berbasis **BERT** dalam menghasilkan ringakasan otomatis tafsir Al-Qur’an dalam dua bahasa, yaitu Arab dan Indonesia. Jenis penelitian ini adalah **penelitian komputasi NLP (Natural Language Processing)** yang bertujuan untuk merancang dan menguji model berbasis **transformer BERT** dalam menggunakan teknik **abstractive summarization**.

3.2 Lokasi dan Waktu Penelitian

3.3 Data dan Sumber Data

Sumber data Utama dalam penelitian ini adalah teks **tafsir Al-Qur'an.** Data diambil dari **korpus publik** tafsir yang sudah tersedia secara online dan bersifat open-access, misalnya dari situs-situs tafsir resmi atau platform digital tafsir yang terpercaya. Data yang diambil meliputi:

a) **Teks Arab tafsir Al-Qur'an**
b) **Terjemahan resmi dalam bahasa Indonesia**

Data ini diolah untuk tujuan summarization dengan proses tokenisasi, segmentasi kalimat, dan penyelarasan antara teks Arab dan Indonesia.

3.4 Teknik Keabsahan Data Preprocessing

Untuk memastikan keabsahan data pada tahap **preprocessing**, beberapa teknik digunakan, antara lain:

a) **Tokenisasi:** Membagi teks menjadi unit-unit kecil (token) untuk memudahkan pemrosesan.
b) **Stopword Removal:** Menghilangkan kata-kata yang dianggap tidak memiliki kontribusi penting, seperti kata sambung atau kata fungsi.
c) **Normalization:** Menormalkan teks dengan menghilangkan karakter khusus, tanda baca, atau simbol yang tidak diperlukan.
d) **Alignment (Penyelarasan):** Menyelaraskan teks Arab dan terjemahannya dalam bahasa Indonesia untuk memastikan kedua bahasa berada dalam satu konteks yang sama.
e) **Cross-validation:** Pengujian dan validasi data melalui pembagian dataset ke dalam beberapa subset, untuk memastikan model dilatih dan diuji pada data yang beragam.

3.5 Teknik Analisis Data

Teknik analisis data yang digunakan adalah evaluasi berbasis metrik **ROUGE** (Recall-Orientef Understudy for Gisting Evaluation), yang merupakan standar untuk mengukur kualitas ringkasan teks. Evaluasi ini mencakup beberapa aspek :

a) **ROUGE-N:** Menghitung jumlah n-grams yang cocok antara ringkasan otomatis dan ringkasan referensi.
b) **ROUGE-L:** Mengukur kesesuaian berdasarkan urutan kalimat terpanjang yang sama.
c) **ROUGE-W:** Menghitung kesamaan berdasarkan urutan kata yang persis sama.

Selain itu, analisis kualitatif dilakukan untuk mengevaluasi seberapa baik ringkasan yang dihasilkan mampu mempertahankan makna dan esensi tafsir Al-Qur'an dalam kedua bahasa. Interpretasi hasil evaluasi mencakup perbandingan

kinerja model hybrid summarization berbasis **BERT** dengan pendekatan summarization lainnya.

Penelitian ini juga mempertimbangkan **kesesuaian semantik** dari ringkasan yang dihasilkan, terutama dalam konteks teks keagamaan yang memerlukan ketelitian lebih dalam menjaga makna teks yang diringkas.

DAFTAR PUSTAKA

Aji Fitra Jaya Institut Perguruan Tinggi Ilmu Al Qur, S. (n.d.). *AL-485 ¶$11DAN HADIS SEBAGAI SUMBER HUKUM ISLAM*.

Alami Merrouni, Z., Frikh, B., & Ouhbi, B. (2023). EXABSUM: a new text summarization approach for generating extractive and abstractive summaries. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00836-y

Alselwi, G., & Taşcı, T. (2024). Extractive Arabic Text Summarization Using PageRank and Word Embedding. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-024-08890-1

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Liu, Y., & Lapata, M. (n.d.). *Text Summarization with Pretrained Encoders*. https://github.com/

Nallapati, R., Zhou, B., dos Santos, C., & Xiang, B. (n.d.). *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*.