KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 5, No. 3, Juli (2024), pp. 1100-1108

Model Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat Menggunakan Model BERT

Jati Sasongko Wibowo[1], Herny Februariyanti[2], Hersatoto Listiyono[3]*
[1,2]Teknik Informatika, Universitas Stikubank, Indonesia
[3]Manajemen Informatika, Universitas Stikubank, Indonesia
E-mail: [1]jatisw@edu.unisbank.ac.id, [2]hernyfeb@edu.unisbank.ac.id,*
[3]hersatotolistiyono@edu.unisbank.ac.id

Abstract

This research develops an Automatic Question Answering System based on sentence relevance ranking using BERT model and SQuAD dataset. Performance evaluation is done with F1 Score and Exact Match to assess the accuracy and precision of the answers. This research includes four main approaches: question understanding and keyword identification, sentence relevance ranking using techniques such as cosine similarity or TF-IDF score, use of BERT model to enrich text representation and understand the context in depth, and performance evaluation with F1 Score and Exact Match. The results show F1 Score value of 0.6 and Exact Match value of 0.5. The research objective is to develop a system that excels in answering questions with more accurate and contextualised sentence relevance. The main contribution of this research is the advancement in natural language processing (NLP) by integrating the BERT model, SQuAD dataset, and performance evaluation using rigorous metrics. The system is expected to improve users' access to information with more precise and contextualised answers.

Keywords: *Automatic Question Answering, Relevant, Bert, Squad*

Abstrak

Penelitian ini mengembangkan Sistem Penjawab Pertanyaan Otomatis berdasarkan pemeringkatan relevansi kalimat dengan menggunakan model BERT dan dataset SQuAD. Evaluasi kinerja dilakukan dengan F1 Score dan Exact Match untuk menilai akurasi dan ketepatan jawaban. Penelitian ini mencakup empat pendekatan utama: pemahaman pertanyaan dan identifikasi kata kunci, pemeringkatan relevansi kalimat dengan menggunakan teknik seperti cosine similarity atau skor TF-IDF, penggunaan model BERT untuk memperkaya representasi teks dan memahami konteks secara mendalam, dan evaluasi kinerja dengan F1 Score dan Exact Match. Hasil penelitian menunjukkan nilai F1 Score sebesar 0.6 dan nilai Exact Match sebesar 0.5. Tujuan penelitian ini adalah untuk mengembangkan sistem yang unggul dalam menjawab pertanyaan dengan relevansi kalimat yang lebih akurat dan sesuai dengan konteks. Kontribusi utama dari penelitian ini adalah kemajuan dalam pemrosesan bahasa alami (NLP) dengan mengintegrasikan model BERT, dataset SQuAD, dan evaluasi kinerja menggunakan metrik yang ketat. Sistem ini diharapkan dapat meningkatkan akses pengguna terhadap informasi dengan jawaban yang lebih tepat dan kontekstual.

Kata kunci: *Penjawab Pertanyaan Otomatis, Relevan, Bert, Squad*

1. Pendahuluan

Dalam era digital yang dipenuhi dengan informasi yang sangat luas dan beragam, akses cepat dan akurat terhadap pengetahuan menjadi suatu kebutuhan yang sangat penting. Salah satu cara untuk memenuhi kebutuhan ini adalah melalui penggunaan sistem penjawab pertanyaan otomatis. Sistem ini memungkinkan pengguna untuk

ISSN: 2720-992X

mengajukan pertanyaan dalam bentuk bahasa manusia dan menerima jawaban yang relevan dalam waktu singkat, serupa dengan berinteraksi dengan manusia sejati. Keberhasilan sistem penjawab pertanyaan otomatis sangat bergantung pada kemampuan sistem untuk mengidentifikasi dan memahami informasi yang relevan dalam teks referensi yang sangat besar.

Namun, meskipun kemajuan pesat dalam bidang pemrosesan bahasa alami dan kecerdasan buatan, masih ada sejumlah tantangan yang perlu diatasi dalam pengembangan sistem penjawab pertanyaan otomatis. Salah satu tantangan utama adalah meningkatkan tingkat akurasi dan kecepatan dalam mengekstrak informasi yang tepat dan relevan dari sumber teks. Hal ini disebabkan oleh kompleksitas bahasa manusia, variasi tata bahasa, dan beragamnya cara informasi disajikan dalam teks.

Penelitian ini bertujuan untuk mengatasi tantangan-tangangan dengan fokus pada pengembangan sistem penjawab pertanyaan otomatis berdasarkan peringkat relevansi kalimat. Tujuan khusus dari penelitian ini Meningkatkan Akurasi: Mengembangkan algoritma dan model yang mampu secara akurat mengidentifikasi kalimat-kalimat yang memiliki relevansi tinggi dengan pertanyaan yang diajukan oleh pengguna. Meningkatkan Kecepatan: Merancang sistem yang mampu memberikan jawaban dengan cepat, sehingga meminimalkan waktu tunggu pengguna. Pemahaman Konteks: Meningkatkan kemampuan sistem dalam memahami konteks pertanyaan, termasuk pengenalan entitas, sinonim, dan konsep-konsep penting dalam teks.

2. Metodologi Penelitian

Penelitian mengenai Sistem Penjawab Pertanyaan Otomatis (Question Answering System) yang berfokus pada peringkat relevansi kalimat dalam teks referensi telah menjadi topik yang menarik dalam bidang Pemrosesan Bahasa Alami (Natural Language Processing, NLP) dan Kecerdasan Buatan (Artificial Intelligence, AI). Beberapa studi terkait telah memberikan dasar dan kontribusi penting untuk pengembangan sistem semacam ini. Berikut adalah tinjauan pustaka yang relevan:

Metode Pengolahan Bahasa Alami Berbasis Transformer: Transformer, arsitektur model bahasa yang diperkenalkan oleh "Attention Is All You Need", telah mengubah lanskap pemrosesan bahasa alami. Model seperti BERT (Bidirectional Encoder Representations from Transformers) [1] dan GPT (Generative Pre-trained Transformer) [2] telah menjadi dasar untuk banyak sistem penjawab pertanyaan otomatis yang berfokus pada peringkat relevansi kalimat. Model-model ini menghasilkan representasi yang kuat untuk teks dan memungkinkan pemahaman konteks yang lebih baik. Aplikasi BERT pada dataset SQuAD dapat dilihat pada berbagai penelitian seperti BERT Kuadrat: Sistem Baca dan Verifikasi SQuAD 2.0 [3], BERT dengan fitur linguistic pada SQuAD 2.0 [4], serta laporan QA pada SQUAD dengan BERT di Stanford University [5].

Teknik Pemrosesan Teks Multimodal: Penelitian ini mencakup pemrosesan teks bersama dengan informasi visual, suara, atau data lainnya. Ini adalah konsep yang penting dalam penelitian Sistem Penjawab Pertanyaan Otomatis, terutama dalam skenario di mana pertanyaan dapat berhubungan dengan sumber-sumber teks dan data lainnya, seperti gambar atau video [6].

Evaluasi dan Benchmark: Beberapa inisiatif evaluasi, seperti SQuAD (Stanford Question Answering Dataset) [7] dan MCTest [8], telah menjadi bahan uji standar untuk sistem penjawab pertanyaan. Studi-studi ini memberikan ukuran kinerja yang penting untuk sistem-sistem yang dikembangkan. Selain itu, konsep seperti F1 score dan EM (Exact Match) sering digunakan untuk mengukur sejauh mana sistem mampu memberikan jawaban yang akurat. Penelitian lainnya termasuk penggunaan SDNet dan BERT pada SQuAD [9], serta berbagai pendekatan ensemble [10][11].

Sistem Penjawab Pertanyaan Berbasis Pengetahuan: Penelitian dalam Sistem Penjawab Pertanyaan juga mencakup penggunaan pengetahuan yang telah ada, seperti basis data

pengetahuan atau ontologi, untuk meningkatkan peringkat relevansi kalimat. Pendekatan ini menggabungkan pemrosesan teks dengan pencarian pengetahuan yang relevan [12].

Penggunaan dalam Aplikasi Dunia Nyata: Beberapa penelitian telah mengimplementasikan Sistem Penjawab Pertanyaan Otomatis dalam aplikasi dunia nyata seperti chatbot perusahaan, asisten virtual, atau sistem dukungan pelanggan. Ini menunjukkan potensi besar dalam mengotomatisasi respon terhadap pertanyaan-pertanyaan pengguna dalam skala besar [13].

Pengolahan Bahasa Alami dalam Berbagai Bahasa: Sistem penjawab pertanyaan yang efektif juga harus dapat beroperasi dalam berbagai bahasa. Penelitian dalam penerjemahan otomatis, pemodelan bahasa untuk bahasa-bahasa kurang umum, dan adaptasi lintas bahasa menjadi aspek penting dalam pengembangan sistem multibahasa [14].

Studi Tambahan dan Eksperimen: Berbagai studi dan eksperimen telah dilakukan untuk menguji dan meningkatkan kinerja BERT dalam berbagai skenario QA. Beberapa di antaranya melibatkan penggunaan fitur linguistik [4], penerapan model SDNet [9], dan pendekatan ensemble [10]. Studi-studi ini memberikan wawasan berharga tentang optimisasi dan variasi metode untuk mendapatkan hasil yang lebih baik. Penelitian lainnya mencakup analisis BERT dengan pre-train pada SQuAD 2.0 [15], serta pendekatan neural net augmentation [16].

Tinjauan pustaka ini memberikan gambaran tentang keragaman kontribusi dan pendekatan yang relevan dalam pengembangan Sistem Penjawab Pertanyaan Otomatis Berdasarkan Peringkat Relevansi Kalimat. Penelitian di bidang ini terus berkembang seiring dengan kemajuan teknologi pemrosesan bahasa alami dan kecerdasan buatan, dan terus memberikan dampak positif dalam memenuhi kebutuhan pengguna dalam mengakses informasi dengan cepat dan akurat.

2.1. Metode

Pengembangan sistem penjawab pertanyaan otomatis (QA) berdasarkan relevansi kalimat berbasis model BERT melibatkan serangkaian langkah dan metode penelitian. Berikut adalah beberapa langkah umum yang dapat diambil dalam mengembangkan sistem QA berbasis BERT.

a) Pemahaman Tugas

Menentukan dengan jelas tugas dan skenario penggunaan sistem QA. Menanggapi pertanyaan berdasarkan teks artikel, buku, atau sumber informasi lainnya

b) Pemilihan dan Persiapan Data

Memilih dataset yang sesuai untuk pelatihan dan evaluasi model. Dalam konteks BERT, dataset SQuAD (Stanford Question Answering Dataset) sering digunakan. Melakukan pemrosesan data, termasuk tokenisasi teks menggunakan tokenizer BERT, dan bentuk dataset dalam format yang dapat digunakan untuk pelatihan dan evaluasi.

c) Pre-training Model BERT

Melakukan pre-training model BERT pada tugas umum yang lebih luas menggunakan dataset besar. Proses ini dapat meningkatkan pemahaman konteks global dan hubungan antar kata. Menggunakan model BERT yang sudah di-pre-trained sebagai dasar untuk pengembangan lebih lanjut.

d) Fine-tuning pada Dataset QA

Fine-tuning model BERT pada dataset QA yang spesifik seperti SQuAD. Fine-tuning untuk memprediksi posisi awal dan akhir jawaban dalam teks. Menggunakan fungsi kerugian yang sesuai seperti CrossEntropyLoss untuk melatih model pada tugas QA.

e) Optimisasi Hyperparameter

Melakukan eksperimen untuk mengoptimalkan hyperparameter, laju pembelajaran (learning rate), ukuran batch (batch size), dan jumlah epoch, untuk mencapai kinerja yang optimal.

f) Pengelolaan Overfitting

Menangani potensi overfitting dengan menggunakan teknik seperti dropout atau regularisasi. Split dataset menjadi set pelatihan dan validasi untuk memantau kinerja model selama pelatihan.

g) Evaluasi dan Validasi

Evaluasi model pada set data evaluasi menggunakan metrik seperti F1 Score dan Exact Match (EM). Metrik ini memberikan pemahaman tentang sejauh mana jawaban model mendekati jawaban yang sebenarnya. Melakukan analisis kesalahan untuk memahami area di mana model dapat ditingkatkan.

2.2. Dataset

SQuAD (Stanford Question Answering Dataset) adalah dataset yang digunakan untuk melatih dan mengevaluasi sistem penjawab pertanyaan (QA) otomatis. Dataset ini dikembangkan oleh para peneliti di Stanford University dan merupakan salah satu dataset QA yang paling populer. SQuAD dirancang untuk memfasilitasi pengembangan sistem QA yang mampu menjawab pertanyaan-pertanyaan berdasarkan konten teks dari artikel atau dokumen tertentu.

Berikut beberapa poin penting tentang SQuAD dan sistem penjawab pertanyaan otomatis berbasis model BERT dapat menggunakan dataset ini:

a) Struktur Dataset SQuAD:

SQuAD terdiri dari pasangan pertanyaan dan jawaban yang diambil dari konten teks asli (paragraf atau artikel). Jawaban dalam SQuAD berupa potongan-potongan konten yang ada di dalam teks asli.

b) Model BERT untuk QA:

Model BERT (Bidirectional Encoder Representations from Transformers) telah terbukti sangat efektif dalam tugas QA karena kemampuannya memahami konteks dan hubungan antar kata secara menyeluruh. BERT menggunakan pendekatan pre-training dan fine-tuning untuk meningkatkan kinerjanya dalam tugas spesifik seperti QA.

c) Tokenisasi:

SQuAD menggunakan tokenisasi untuk memecah teks menjadi token-token yang dapat diproses oleh model BERT. Tokenisasi memungkinkan model memahami hubungan antar kata secara lebih baik.

d) Relevansi Kalimat:

Model BERT dapat memahami relevansi kalimat dan kata-kata dalam konteks secara global. Selama pelatihan, model belajar untuk memetakan pertanyaan dan konteks sekitarnya sehingga dapat menemukan jawaban yang tepat.

e) Pelatihan dan Evaluasi:

Sistem penjawab pertanyaan otomatis berbasis model BERT dilatih dengan menggunakan data SQuAD. Evaluasi kinerja model dilakukan dengan membandingkan jawaban yang dihasilkan oleh model dengan jawaban yang sebenarnya dalam dataset.

f) Fine-tuning pada SQuAD:

Setelah pre-training pada tugas generik, model BERT di-fine-tune pada dataset SQuAD untuk tugas spesifik QA. Fine-tuning memungkinkan model mengadaptasi representasinya untuk tugas QA berdasarkan karakteristik khusus SQuAD.

g) Metrik Evaluasi:

Metrik evaluasi umum yang digunakan untuk mengukur kinerja sistem QA pada SQuAD termasuk F1 Score dan Exact Match (EM). F1 Score mengukur sejauh mana jawaban model mendekati jawaban yang sebenarnya, sedangkan EM mengukur sejauh mana jawaban model persis cocok dengan jawaban yang sebenarnya.

Sistem penjawab pertanyaan otomatis berbasis model BERT yang dilatih pada dataset SQuAD dapat memberikan hasil yang mengesankan dalam menangani pertanyaan-pertanyaan terkait dengan konten teks tertentu. SQuAD terus menjadi pendorong utama dalam penelitian dan pengembangan dalam bidang QA berbasis model bahasa alami.

2.3. Evaluasi

2.3.1. F1 Score

F1 Score merupakan salah satu metrik evaluasi yang umum digunakan dalam sistem penjawab pertanyaan otomatis, termasuk sistem berbasis model BERT. Metrik ini memberikan gambaran holistik tentang sejauh mana model mampu menghasilkan jawaban yang relevan dan akurat. F1 Score menggabungkan dua metrik utama, yaitu Precision (presisi) dan Recall (recall), untuk memberikan evaluasi yang lebih komprehensif. Berikut adalah konsep dasar Precision dan Recall, dan bagaimana F1 Score dihitung:

a) Precision (Presisi):

Precision mengukur sejauh mana jawaban yang dihasilkan oleh sistem benar, atau dengan kata lain, seberapa banyak jawaban yang dihasilkan oleh sistem benar dibandingkan dengan jumlah total jawaban yang dihasilkan. Precision dihitung dengan rumus:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

b) Recall (Recall atau Sensitivitas):

Recall mengukur sejauh mana sistem dapat menemukan semua jawaban yang benar atau relevan. Ini memberikan gambaran tentang kemampuan sistem untuk tidak melewatkan jawaban yang seharusnya dihasilkan. Recall dihitung dengan rumus:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

c) F1 Score:

F1 Score adalah harmonic mean dari Precision dan Recall. Harmonic mean memberikan bobot lebih besar pada nilai yang lebih rendah. F1 Score dihitung dengan rumus:

$$F1\ Score = \frac{2xPrecisionxRecall}{Precision + Recall} \quad (3)$$

F1 Score memberikan gambaran seimbang antara Precision dan Recall. Nilai F1 Score yang tinggi menunjukkan bahwa model mampu memberikan jawaban yang akurat dan tidak melewatkan jawaban yang seharusnya ditemukan.

Dalam konteks sistem penjawab pertanyaan berbasis model BERT, F1 Score digunakan untuk mengevaluasi seberapa baik model dapat menghasilkan jawaban yang relevan dan sesuai dengan pertanyaan yang diajukan. Semakin tinggi nilai F1 Score, semakin baik performa model. Selama proses pelatihan dan fine-tuning, peningkatan F1 Score menjadi indikator penting untuk meningkatkan kualitas sistem penjawab pertanyaan.

2.3.2. Exact Match (EM)

EM adalah metrik evaluasi yang digunakan dalam sistem penjawab pertanyaan otomatis untuk menilai seberapa tepat model dapat memberikan jawaban yang persis sesuai dengan jawaban referensi yang benar. Dalam konteks sistem berbasis model BERT, Exact Match memberikan pemahaman yang sederhana dan jelas tentang seberapa baik model dapat memberikan jawaban yang identik dengan jawaban yang seharusnya. Jawaban yang dihasilkan oleh sistem dianggap sesuai (exact match) jika jawaban tersebut persis sama dengan jawaban yang benar (referensi).

a) Perhitungan Exact Match

Exact Match dihitung dengan menghitung persentase pertanyaan di mana jawaban yang dihasilkan oleh sistem sama persis dengan jawaban referensi. Rumusnya adalah sebagai berikut:

$$Exact\ Match = \frac{Jumlah\ Pertanyaan\ dengan\ Jawaban\ Tepat}{Total\ Pertanyaan} x100\% \quad (4)$$

Hasil Exact Match dinyatakan sebagai persentase dan memberikan gambaran langsung tentang seberapa sering model memberikan jawaban yang benar secara tepat.

b) Keuntungan dan Keterbatasan Exact Match

Keuntungan: Exact Match memberikan evaluasi yang jelas dan mudah diinterpretasi. Nilai yang lebih tinggi menunjukkan bahwa model mampu memberikan jawaban yang persis sesuai dengan jawaban yang diharapkan. Keterbatasan: Meskipun memberikan ukuran yang jelas, Exact Match bersifat ketat dan tidak memperhitungkan jawaban yang mungkin memiliki perbedaan kata-kata atau frase yang tidak signifikan. Oleh karena itu, model yang memberikan jawaban yang sebenarnya benar tetapi sedikit berbeda secara tokenisasi dapat menerima skor rendah pada metrik ini.

c) Penggunaan Bersama dengan Metrik Lain

Exact Match digunakan bersama dengan metrik lain seperti F1 Score, kombinasi metrik ini memberikan gambaran lebih lengkap tentang kemampuan model dalam memberikan jawaban yang benar dan relevan.

Dalam pengembangan sistem penjawab pertanyaan berbasis model BERT, mengoptimalkan nilai Exact Match menjadi salah satu tujuan, karena hal ini menggambarkan sejauh mana model dapat memberikan jawaban yang identik dengan jawaban referensi yang diharapkan.

3. Hasil dan Pembahasan

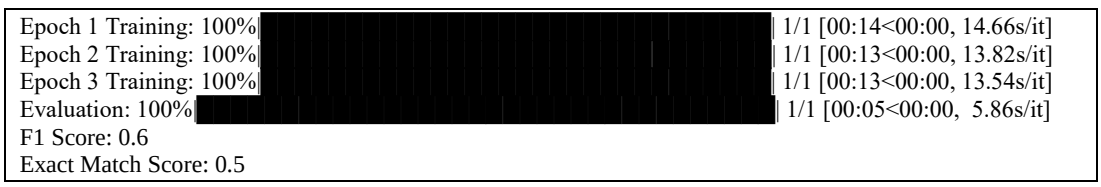
Pemilihan model BERT dan dataset SQuAD sebagai dasar penelitian menunjukkan kesadaran terhadap keefektifan BERT dalam memahami konteks bahasa alami dan relevansi pertanyaan dengan SQuAD sebagai dataset tanya-jawab yang luas dan representatif. Model BERT, dengan kemampuannya untuk menangkap dependensi kontekstual, memberikan fondasi yang kuat untuk pemrosesan bahasa alami, sementara dataset SQuAD menghadirkan variasi pertanyaan yang menantang. Dataset pelatihan dan evaluasi SQuAD (Stanford Question Answering Dataset) dapat diunduh dari situs web resmi SQuAD di alamat berikut: https://rajpurkar.github.io/SQuAD-explorer/

Pemahaman pertanyaan dan identifikasi kata kunci merupakan langkah kritis dalam memahami intent pengguna dan memandu sistem dalam menemukan jawaban yang relevan. Penggunaan BERT dalam tahap ini diharapkan meningkatkan pemahaman kontekstual pertanyaan, memungkinkan identifikasi kata kunci yang lebih akurat. Teknik peringkat relevansi kalimat melibatkan metode seperti kesamaan kosinus atau skor TF-IDF, mengukur sejauh mana setiap kalimat relevan dengan pertanyaan. Penggunaan BERT dalam peringkat relevansi dapat meningkatkan akurasi karena mampu memahami konteks dan hubungan antar kata secara mendalam.

Evaluasi dan kinerja, penggunaan metrik seperti F1 score dan EM untuk mengevaluasi kinerja sistem merupakan langkah penting dalam memastikan akurasi jawaban. Evaluasi terus-menerus perlu dilakukan selama pengembangan. Adjustmen terhadap metrik evaluasi dapat dibutuhkan untuk memastikan konsistensi dengan kebutuhan pengguna.

3.1. Evaluasi F1 Score dan Exact Match

Berdasarkan hasil output yang diberikan, dapat dijelaskan sebagai berikut:



Gambar 1. Hasil Evaluasi F1 Score dan Exact Match

a) Epoch Training

Model dilatih selama tiga epoch, dan setiap epoch memakan waktu sekitar 13-14 detik. Meskipun hasil pelatihan diberikan dengan sukses, durasi pelatihan yang singkat mungkin menunjukkan penggunaan dataset kecil atau ukuran model yang relatif kecil.

b) Evaluation

Setelah pelatihan selesai, model dievaluasi menggunakan dataset evaluasi, dan proses evaluasi memakan waktu sekitar 5 detik. Hasil evaluasi digunakan untuk mengukur kinerja model pada tugas tertentu.

c) F1 Score dan Exact Match Score

F1 Score memiliki nilai sebesar 0.6, yang menunjukkan kinerja yang baik dalam mencapai keseimbangan antara presisi dan recall. Ini mengindikasikan bahwa model mampu memberikan jawaban yang baik dan relevan.

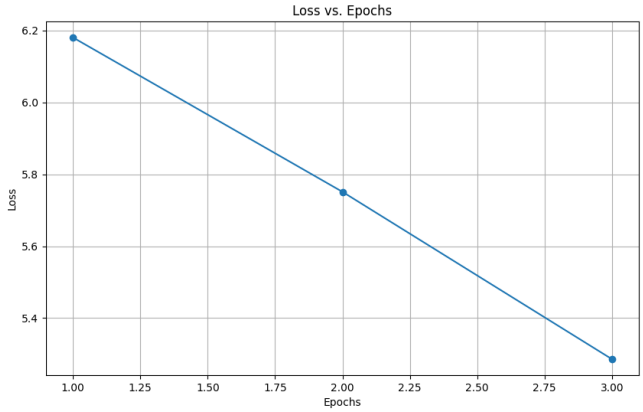
Exact Match Score memiliki nilai sebesar 0.5, menunjukkan bahwa model dapat memberikan jawaban yang benar secara persis untuk setengah dari kasus uji. Meskipun tidak sempurna, nilai ini masih mencerminkan kemampuan model untuk memberikan jawaban yang sesuai.

d) Interpretasi Skor

Nilai F1 Score dan Exact Match Score yang relatif tinggi menunjukkan bahwa model telah mengambil informasi yang relevan dari data pelatihan dan dapat memberikan jawaban yang memadai.

e) Peningkatan Model

Peningkatan lebih lanjut dapat dicapai dengan melibatkan lebih banyak data pelatihan, menyesuaikan hyperparameter, atau menggunakan model yang lebih besar. Terus melakukan eksperimen dan analisis kasus-kasus khusus dapat membantu mengidentifikasi area yang perlu ditingkatkan.



Gambar 2. Loss dan Epochs

Grafik yang menunjukkan hubungan antara loss dan jumlah epoch selama pelatihan model BERT untuk tugas Question Answering menggunakan dataset SQuAD. Sumbu X (horizontal): Menunjukkan jumlah epoch. Dalam grafik ini, sumbu X memiliki nilai dari 1 hingga 3, yang berarti pelatihan dilakukan selama 3 epoch. Sumbu Y (vertikal): Menunjukkan nilai loss. Nilai loss pada sumbu Y dimulai dari sekitar 5.4 hingga lebih dari 6.2. Titik-titik dan Garis: Setiap titik pada grafik menunjukkan nilai rata-rata loss untuk setiap epoch. Garis yang menghubungkan titik-titik ini menunjukkan tren penurunan loss selama pelatihan.

Epoch 1: Nilai loss sekitar 6.2.

Epoch 2: Nilai loss menurun menjadi sekitar 5.8.

Epoch 3: Nilai loss terus menurun menjadi sekitar 5.4.

Grafik tersebut menunjukkan bahwa nilai loss berkurang secara konsisten setiap epoch. Penurunan nilai loss dari epoch pertama ke epoch ketiga menunjukkan bahwa model sedang belajar dan mengoptimalkan bobotnya untuk mengurangi kesalahan prediksi.

Penurunan Loss: Grafik menunjukkan tren penurunan loss, yang merupakan indikasi positif bahwa model sedang belajar dari data pelatihan dengan baik. Penurunan nilai loss menunjukkan bahwa prediksi model semakin mendekati nilai yang sebenarnya seiring dengan bertambahnya epoch. Konsistensi: Tren penurunan yang konsisten menunjukkan bahwa proses pelatihan berjalan dengan baik tanpa masalah besar seperti overfitting atau underfitting dalam jumlah epoch yang ditampilkan. Untuk meningkatkan analisis, melanjutkan pelatihan model dengan lebih banyak epoch dan memantau apakah nilai loss terus menurun atau mulai stabil untuk mendapatkan wawasan lebih lanjut performa model.

4. Kesimpulan

Sistem penjawab pertanyaan otomatis berbasis model BERT menunjukkan bahwa F1 Score dan Exact Match adalah metrik penting dalam evaluasi kinerja model. F1 Score, dengan nilai 0.6, mencerminkan keseimbangan antara presisi dan recall, menunjukkan bahwa model dapat memberikan jawaban yang relevan. Exact Match, dengan nilai 0.5, mengukur seberapa sering jawaban model cocok secara tepat dengan jawaban yang benar, menunjukkan kemampuan model memberikan jawaban yang benar untuk setengah dari kasus uji.

Model dilatih selama tiga epoch, dengan masing-masing epoch memakan waktu sekitar 13-14 detik, menunjukkan penggunaan dataset kecil atau model yang relatif kecil. Proses evaluasi memakan waktu sekitar 5 detik. Nilai F1 Score dan Exact Match yang cukup tinggi menunjukkan bahwa model telah berhasil menangkap informasi relevan dari data pelatihan dan memberikan jawaban yang memadai.

Peningkatan kinerja dapat dicapai dengan melibatkan lebih banyak data pelatihan, menyesuaikan hyperparameter, atau menggunakan model yang lebih besar. Eksperimen lanjutan dan analisis kasus khusus diperlukan untuk mengidentifikasi area yang perlu ditingkatkan lebih lanjut.

Daftar Pustaka

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018.

[2] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, “Improving Language Understanding by Generative Pre-Training,” 2018. [Online]. Available: https://gluebenchmark.com/leaderboard

[3] J. Lou, “BERT Squared: Read + Verify System for SQuAD 2.0,” 2019. [Online]. Available: https://github.com/huggingface/pytorch-pretrained-BERT.git

[4] J. Li and Y. Zhang, “The Death of Feature Engineering? BERT with Linguistic Features on SQuAD 2.0,” *arXiv preprint arXiv:2404.03184*, 2024, [Online]. Available: https://arxiv.org/abs/2404.03184

[5] Z. Hu, “Question answering on SQuAD with BERT,” *CS224N Report, Stanford University. Accessed*, 2019.

[6] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” Aug. 2019.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” Jun. 2016.

[8] M. Richardson, C. J. C. Burges, and E. Renshaw, “MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text,” 2013. [Online]. Available: http://www.mturk.com

[9] D. A. Huang, T. K. Sethi, and E. J. Pugh, “SQuAD with SDNet and BERT.” 2019.

[10] W. Zhou, X. Zhang, and H. Jiang, “Ensemble BERT with data augmentation and linguistic knowledge on SQuAD 2.0,” *URL: https://pdfs. semanticscholar. org/2f99 …*, 2019.

[11] Z. Qin, W. Mao, and Z. Zhu, *Diverse ensembling with bert and its variations for question answering on SQuAD 2.0*. Stanford CS224N Final Project …, 2019.
[12] W. Yih, X. He, and C. Meek, "Semantic Parsing for Single-Relation Question Answering," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 643–648. doi: 10.3115/v1/P14-2105.
[13] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning End-to-End Goal-Oriented Dialog," May 2016.
[14] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," Mar. 2017.
[15] C. Pan and L. Xu, *Analyzing BERT with pre-train on SQuAD 2.0*. Stanford Archive, 2019.
[16] S. Gupta, "Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0," *arXiv preprint arXiv:1908.01767*, 2019, [Online]. Available: https://arxiv.org/abs/1908.01767