Prosiding Seminar Nasional Sains dan Teknologi Seri III
Fakultas Sains dan Teknologi, Universitas Terbuka
Vol. 2 No. 1 (2025)
e-ISSN: 3047-6569

IMPLEMENTASI SENTENCE TRANSFORMER BERBASIS BERT UNTUK EDUKASI DAN INFORMASI TENTANG HIV/AIDS

Martin Ompusunggu[1*], A Rafi Paringgom Iwari[2], Silvia Azahrani[2], Devina Anin Safitri[2], Annisa Cahyani Surya[2], Nadinda Aurora[2], Ardika Satria[2], Christyan Tamaro Nadeak[2], Ade Lailani[2]

[1]Teknik Informatika, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
[2]Sains Data, Institut Teknologi Sumatera, Lampung Selatan, Indonesia

**Penulis korespondensi: martin.121140144@student.itera.ac.id*

ABSTRAK

Human Immunodeficiency Virus (HIV) atau *Acquired Immunodeficiency Syndrome* (AIDS) menjadi salah satu masalah kesehatan global yang memerlukan perhatian serius, khususnya pada kelompok usia produktif. Dalam upaya meningkatkan kesadaran dan pemahaman masyarakat, teknologi *Artificial Intelligence* (AI) dan *Deep Learning* dapat dimanfaatkan sebagai media edukasi yang inovatif. Penelitian ini bertujuan untuk mengembangkan chatbot *Sentence Transformer* berbasis *Bidirectional Encoder Representations from Transformers* (BERT) yang mampu memberikan respons relevan terhadap pertanyaan pengguna seputar HIV/AIDS. Dalam prosesnya, penelitian ini melibatkan pengolahan dataset pasangan pertanyaan dan jawaban terkait HIV/AIDS, pemanfaatan *semantic embedding*, dan menggunakan *cosine similarity* untuk pengukuran kesamaan. Dalam melatih model dataset akan diolah untuk menghasilkan respon *chatbot* yang menghasilkan nilai kesamaan tertinggi melebihi ambang batas yang dijaga. Akurasi performa *chatbot* dievaluasi menggunakan dataset uji. Hasil penelitian menunjukan bahwa *chatbot* memiliki rata-rata akurasi jawaban dari data uji sebesar 92% yang menunjukan model *Transformer* berbasis BERT memiliki potensi untuk mendukung edukasi mengenai HIV/AIDS. Penelitian ini memberikan kontribusi awal dalam pengaplikasian teknologi AI untuk mendukung edukasi kesehatan masyarakat.

Kata kunci: BERT, *chatbot*, edukasi kesehatan, HIV/AIDS, *Sentence Transformer*.

1 PENDAHULUAN

Perkembangan teknologi informasi dan kecerdasan buatan telah membuka peluang transformatif dalam diseminasi informasi kesehatan, salah satunya dalam konteks HIV/AIDS. HIV (*Human Immunodeficiency Virus*) adalah virus yang menyerang sistem kekebalan tubuh yang dapat melemahkan kemampuan tubuh melawan infeksi dan penyakit. AIDS (*Acquired Immune Deficiency Syndrome*) adalah kondisi di mana HIV sudah pada tahap infeksi akhir (Purba et al., 2024, 1). Menurut data Kementerian Kesehatan Indonesia ketika tahun 2021, tercatat 36.902 kasus HIV. Kebanyakan pengidap HIV ini berusia pada rentang umur produktif. Kasus HIV paling banyak terjadi pada kelompok usia 25-49 tahun, mencapai 69,7%, diikuti oleh kelompok umur 20-24 tahun 16,9%, serta kelompok umur 15-19 tahun sebanyak 3,1%. Selain itu, total kasus AIDS atau *Acquired Immune Deficiency Syndrome* pada Indonesia saat tahun 2021 juga kebanyakan terjadi untuk umur 30-39 tahun (Maulana et al., 2024, 1).

HIV/AIDS merupakan salah satu masalah kesehatan global yang memerlukan perhatian serius,

baik dari segi pencegahan, pengobatan, maupun edukasi masyarakat. Penyakit ini tidak hanya berdampak pada kesehatan individu tetapi juga memiliki konsekuensi sosial, ekonomi, dan budaya yang luas. Dalam upaya meningkatkan kesadaran dan pemahaman masyarakat, pemanfaatan teknologi informasi menawarkan peluang besar untuk menyediakan informasi yang relevan dan berbasis data. Laporan ini berfokus pada penerapan model *Transformer*, khususnya *Sentence Transformer* (BERT). Perkembangan yang paling berpengaruh terhadap bidang kecerdasan buatan adalah dengan diperkenalkannya arsitektur *Transformer*. *Transformer* merupakan sebuah arsitektur model kecerdasan buatan yang pertama kali diusulkan pada 2017, untuk menggantikan penggunaan pendekatan sekuensial yang digunakan dalam model-model kecerdasan buatan yang telah banyak dilakukan sebelumnya, antara lain pendekatan dengan menggunakan arsitektur *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM) (Firmanto et al., 2024, 1).

Melalui pendekatan semantik berbasis embedding, model yang diusulkan bertujuan menghasilkan respon yang relevan dan mendalam terhadap berbagai pertanyaan terkait HIV/AIDS. Hal ini memanfaatkan kemampuan model bahasa mutakhir untuk menginterpretasikan konteks dan menghasilkan jawaban yang akurat, mencakup spektrum luas topik mulai dari informasi dasar, pencegahan, diagnostik, hingga aspek sosial dan medis yang kompleks.

Tujuan penelitian ini adalah untuk mengembangkan chatbot berbasis Bidirectional Encoder Representations from Transformers (BERT) yang mampu memberikan respons relevan terhadap pertanyaan terkait HIV/AIDS. Selain itu, penelitian ini juga bertujuan mengevaluasi akurasi dan efektivitas chatbot sebagai media edukasi untuk mendukung kesehatan masyarakat. Adapun manfaat yang diharapkan dari penelitian ini meliputi penyediaan media edukasi yang inovatif untuk meningkatkan kesadaran dan pemahaman masyarakat tentang HIV/AIDS, serta mempermudah akses informasi kesehatan yang dapat diakses kapan saja dan di mana saja.

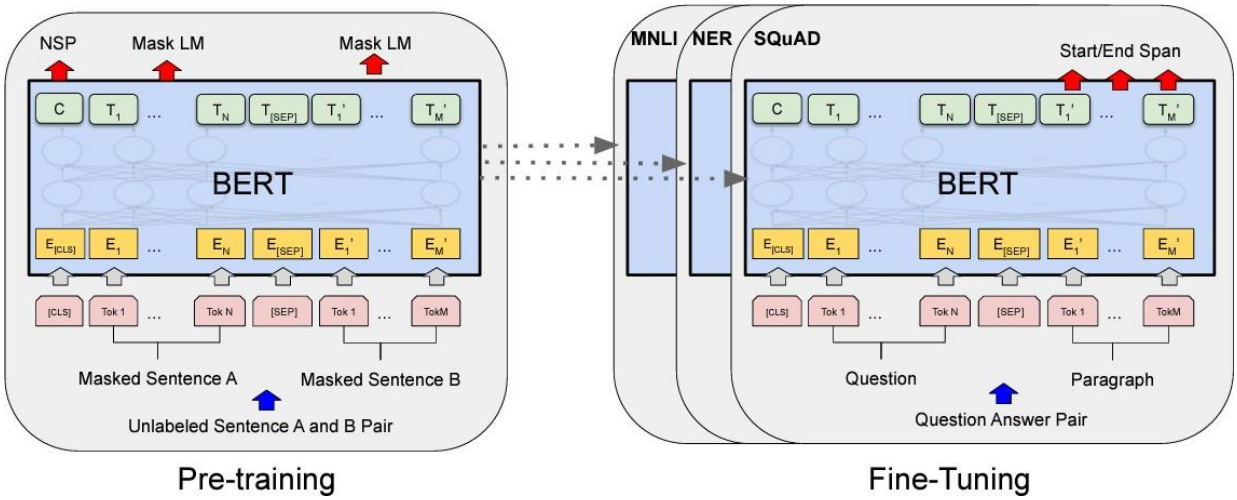
2 METODE

2.1 Data

Data yang digunakan dalam membangun model dengan asitektur *sentence transformer* berbasis BERT akan dibuat dataset dengan format json, yang memiliki struktur *intent,* dimana berbasis *key-value,* key dalam data diantaranya tujuan, pertanyaan dan jawaban. Setiap *key* diisi oleh konten yang berasal dari buku-buku referensi yang membahasa mengenai HIV/AIDS atau sejenisnya. Melalui rencangan ini, model *sentence transformer* berbasis BERT untuk menuju program *chatbot* edukasi HIV dapat lebih mudah diimplementasikan.

2.2 Bert

Penelitian ini mengimplementasikan BERT yang merupakan cabang dari *Sentence Transformer*. Terdapat dua langkah pada BERT yaitu *Pre-training* dan *Fine-Tuning.* Namun kami hanya menggunakan *Pre-training* yang akan langsung di implementasikan ke model utama. Selama *Pre-training,* model dilatih dengan data yang akan menghasilkan parameter dan *hyperparameter* yang digunakan untuk proses *training* pada model utama, sehingga model memahami struktur bahasa dan hubungan antar kata sebagai kemampuan dasar.



Gambar 1. Diagram proses *Pre-training* dan *Fine-Tuning*

Bidirectional Encoder Representations from Transformers (BERT) memanfaatkan arsitektur *Sentence Transformer* untuk membantu model memahami konteks dalam teks secara mendalam. Dengan pendekatan *bidirectional*, BERT membaca konteks dari arah maju dan mundur, memungkinkan pemahaman yang lebih kaya terhadap makna kata dalam kalimat. Hal ini mengatasi kelemahan model sebelumnya yang sering menghasilkan representasi global tetap untuk suatu kata, terlepas dari konteksnya. BERT memanfaatkan *self-attention* untuk menghitung hubungan antara setiap token dalam teks, sehingga setiap kata dipengaruhi oleh seluruh kata lain di sekitarnya (Aljabar & Karomah, 2024).

Dalam arsitektur *Transformer*, setiap input teks melalui serangkaian proses *encoder*, yang bertugas memahami hubungan dan representasi kontekstual dari teks. *Encoder* bekerja melalui mekanisme *multi-head attention* dan *feed-forward neural network*, menghasilkan representasi yang memuat informasi untuk setiap token. Untuk mencegah kebocoran informasi masa depan selama pelatihan, digunakan metode *Masked Language Modeling* (MLM), di mana sebagian token dalam masukkan disembunyikan dan model dilatih untuk memprediksi token yang hilang. Sementara itu, *decoder*, yang biasanya digunakan pada *Transformer* untuk menghasilkan keluaran seperti terjemahan, tidak menjadi bagian utama dalam BERT karena fokus utamanya adalah pemahaman teks, bukan generasi teks. Arsitektur ini menjadikan BERT sangat efektif untuk berbagai tugas pemrosesan bahasa alami, seperti klasifikasi teks, analisis sentimen, dan pencocokan pertanyaan-jawaban.

2.3 *Cosine Similarity*

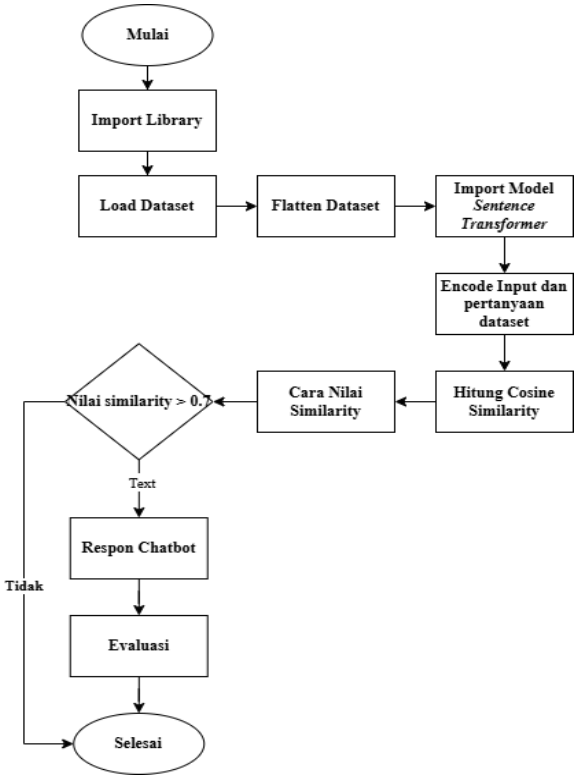
Pengukuran terhadap performa model *text generation* digunakan proses perhitungan dengan metode relevansi *cosine similarity*, yang akan menghitung *score* kesamaan antara respon atau jawaban yang dihasilkan oleh model dengan jawaban yang ada dalam format *intent* dengan *threshold* 70%, sehingga skor kesamaan antara respon dan jawaban dalam format *intent*, sehingga memastikan keluaran yang diberikan oleh model akan relevan melalui persamaan.

$$\textbf{Cosine Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

dengan $\mathbf{A}\cdot\mathbf{B}$ adalah *dot product* dari dua vektor **A** dan **B**, dihitung sebagai penjumlahan perkalian setiap vektor $\boldsymbol{A} \cdot \boldsymbol{B} = \sum_{i=1}^{n} A_i B_i$, serta norm (panjang) vektor akan dihitung sebagai penjumlahan kuadrat yang diakarkan yang dinyatakan sebagai, dan $||B|| = \sqrt{B^2}$. Evaluasi dengan *cosine similarity* ini memberikan skor kesamaan antara 0 hingga 1, di mana skor mendekati 1 menunjukkan tingkat kesamaan yang tinggi antara respon model dan jawaban referensi dalam format *intent*. Dengan *threshold* 70%, jika skor kesamaan lebih besar atau sama dengan 0.7, maka keluaran model dianggap relevan. Evaluasi ini memastikan model tidak hanya menghasilkan teks yang secara sintaksis benar tetapi juga kontekstual dan sesuai dengan tujuan yang diharapkan.

2.4 Diagram Alir Penelitian

Penelitian ini dimulai dengan memasukkan dataset yang berisi pasangan pertanyaan dan jawaban. Selanjutnya, data tersebut diproses melalui *flatten* untuk memastikan formatnya sesuai dengan kebutuhan model. Model *Text Generation* dari *Hugging Face* kemudian di *import*, dan masukan pengguna serta seluruh pertanyaan dalam dataset di *encode* menggunakan representasi vektor. Kemudian menghitung *Cosine Similarity* antara masukan pengguna dan pertanyaan dalam dataset untuk mengukur tingkat kesamaan. Berdasarkan nilai *similarity*, ditentukan *threshold* sebagai batas pengambilan keputusan. Pertanyaan dengan nilai *similarity* tertinggi dipilih, dan respon yang sesuai ditampilkan kepada pengguna. Akhirnya, model dievaluasi untuk menilai kinerjanya dalam menghasilkan respon yang relevan. Langkah penelitian yang sudah dibentuk dalam diagram alir penelitian dapat dilihat pada **Gambar 2.**



Gambar 2. Diagram Alir Penelitian

3 HASIL DAN PEMBAHASAN

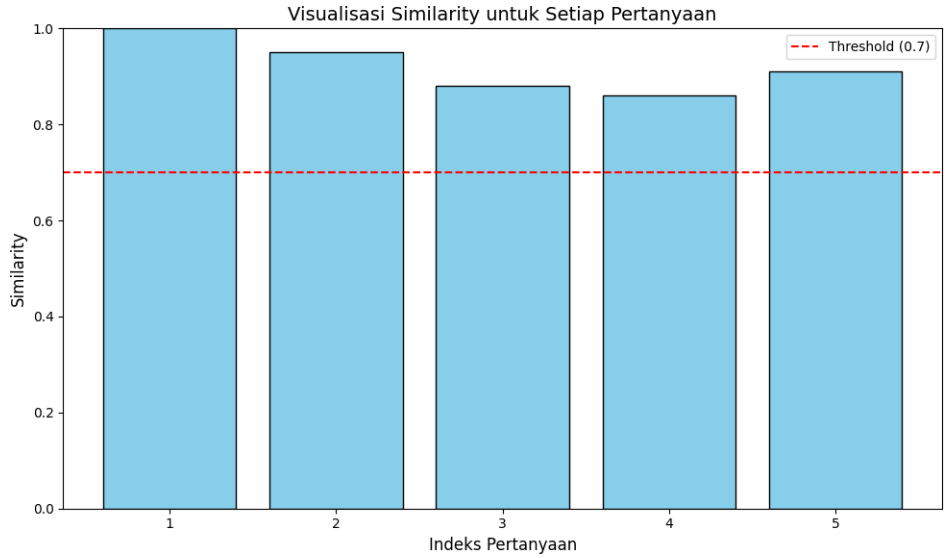
Berdasarkan penerapan model *Sentence Transformer* menggunakan arsitektur BERT pada data informasi mengenai HIV/AIDS, model berhasil memberikan respons yang relevan terhadap pertanyaan yang diberikan oleh pengguna. Dataset yang mencakup berbagai kategori informasi dalam bentuk pertanyaan dan jawaban mengenai HIV/AIDS digunakan dalam melatih model. Setelah pelatihan dilakukan selama 3 epoch dengan masing-masing memiliki *batch size* sebesar 32 dan *oversampling* untuk menangani ketidakseimbangan data, model dapat memberikan respons yang cukup tepat berdasarkan *Cosine Similarity.*

Untuk mengevaluasi kualitas respons yang diberikan oleh *chatbot*, digunakan beberapa metrik evaluasi, yaitu *Precision*, *Recall*, dan *F1-Score*. Ketiga metrik ini digunakan untuk mengukur performa model, sehingga dihasilkan nilai akurasi dari model *Sentence Transformer* yang diterapkan. Proses evaluasi dilakukan dengan menguji model menggunakan beberapa percobaan masukkan berupa pertanyaan yang disesuaikan dengan jawaban dari dataset yang digunakan. Dalam proses ini, pertanyaan yang dimasukkan ke dalam chatbot dirancang agar memiliki inti yang serupa dengan pertanyaan dalam dataset. Jawaban yang diharapkan (*expected output*) merujuk pada jawaban yang telah ditentukan dalam dataset untuk setiap pertanyaan. Selanjutnya, model akan memberikan respons berdasarkan pemrosesan input tersebut. Hasil respons dari model kemudian dibandingkan dengan *expected output* untuk menilai sejauh mana *chatbot* mampu memberikan jawaban yang akurat dan relevan. Evaluasi ini dilakukan untuk memastikan bahwa model dapat memberikan respons yang mendekati atau sesuai dengan yang diharapkan, sehingga memenuhi kebutuhan pengguna secara optimal. **Tabel 2.** menunjukkan hasil percobaan pengujian *chatbot* menggunakan model *Sentence Transformer*, termasuk *input*, jawaban yang diharapkan, respon *chatbot*, dan tingkat kemiripan (*similarity*) yang dihasilkan.

Tabel 2. Percobaan pengujian *Chatbot* dengan model *Sentence Transformer*

Input	*Expected*	*Chatbot*	*Similarity*
Jelaskan apa itu HIV?	Singkatan dari *Human Immunodeficiency Virus*, sebuah virus yang menyerang sistem kekebalan tubuh manusia.	HIV adalah virus yang menyerang sistem kekebalan tubuh.	1.00
Bagaimana cara penularan HIV?	Hubungan Seks Tanpa Alat pengaman, Berbagi Alat Suntik, Ibu Hamil ke Bayi, Transfusi Darah, Seks Oral, Pertukaran Cairan Vagina atau Sperma, Ibu yang terinfeksi HIV dan menyusui dapat menularkan virus kepada bayinya melalui ASI.	HIV dapat menular melalui kontak dengan darah, hubungan seksual, atau dari ibu ke anak selama kehamilan.	0.95

Bisakah HIV disembuhkan?	Sampai saat ini, belum ada obat yang dapat menyembuhkan HIV. Namun, ada obat antiretroviral (ARV) yang dapat memperlambat perkembangan virus dan membantu menjaga sistem kekebalan tubuh.	Saat ini belum ada obat untuk menyembuhkan HIV, tetapi pengobatan antiretroviral (ARV) dapat mengontrol virus.	0.86
Apa gejala awal HIV?	Tanda-tanda awal infeksi HIV meliputi demam yang disertai ruam makulopapular, ulserasi mukosa (di mulut, anus atau penis), sindrom retroviral akut yang menyerupai flu, serta gejala neurologis seperti meningitis aseptik.	Gejala awal HIV mirip dengan Flu, seperti demam, sakit kepala, dan ruam kulit.	0.88
Bagaimana cara mencegah HIV?	Hubungan Seks Tanpa Alat pengaman, Berbagi Alat Suntik, Ibu Hamil ke Bayi, Transfusi Darah, Seks Oral, Pertukaran Cairan Vagina atau Sperma, Ibu yang terinfeksi HIV dan menyusui dapat menularkan virus kepada bayinya melalui ASI.	Menggunakan kondom, tidak berbagi jarum suntik, dan melakukan tes HIV secara rutin dapat mencegah penularan HIV.	0.91



Gambar 3. *Similarity* untuk setiap pertanyaan

Gambar 3. menunjukkan *similarity* untuk setiap pertanyaan. Berdasarkan hasil evaluasi model *chatbot* menggunakan *Sentence Transformer*, visualisasi diagram batang menunjukkan tingkat similarity dari lima data uji berturut-turut sebesar 100%, 95%, 86%, 88%, dan 91%. Performa model menunjukkan rata-rata akurasi jawaban sebesar 92%, yang berarti model mampu memberikan jawaban dengan tingkat kesamaan yang baik untuk 92% dari total data uji. *Precision* yang dihasilkan sebesar 100%, mengindikasikan bahwa secara keseluruhan respon yang diberikan model relevan. Sementara itu, *Recall* sebesar 100% menunjukkan bahwa model berhasil menangkap mayoritas respon yang seharusnya diidentifikasi dengan benar. Kombinasi antara *Precision* dan *Recall* menghasilkan *F1-Score* sebesar 100%, yang mencerminkan performa model pada tingkat yang moderat.

Dalam proses evaluasi, peneliti menggunakan hyperparameter berupa *threshold* sebesar 0.7 untuk menentukan kesesuaian antara respons dan jawaban. Secara umum, model ini bekerja dengan cukup baik dalam memahami dan menjawab pertanyaan dari data uji. Pengembangan lebih lanjut, seperti penambahan data latih dan pengoptimalan model, dapat membantu meningkatkan performa secara keseluruhan agar lebih sesuai dengan kebutuhan pengguna.

4 KESIMPULAN

Berdasarkan hasil yang didapat, pengembangan *chatbot* berbasis *Bidirectional Encoder Representations from Transformers* (BERT), menggunakan arsitektur dari *Sentence Transformer* dengan model BERT bernama *Distiluse-Base-Multilingual-Cased-V2* berhasil memberikan respons yang relevan atas pertanyaan terkait HIV/AIDS. Model *chatbot* ini dilatih menggunakan dataset yang mencakup beberapa kategori informasi mengenai HIV/AIDS, serta menggunakan metrik akurasi, *Precision*, *Recall*, dan *F1-Score* untuk evaluasi. Hasil evaluasi menunjukan performa yang baik yakni kesamaan kata 92%, *Precision* 100%, *Recall* 100%, dan *F1-Score* 100%, menunjukan bahwa performa cukup baik untuk memberikan jawaban yang relevan meskipun masih memiliki ruang untuk pengembangan. Sehingga, dari hasil yang diperoleh chatbot dapat menjadi salah satu media edukasi kesehatan masyarakat yang efektif, meskipun pengembangan lebih lanjut perlu dilakukan untuk meningkatkan kinerjanya.

UCAPAN TERIMAKASIH

Ucapan terimakasih disampaikan kepada para dosen tim pengajar mata kuliah *Deep Learning* Program Studi Sains Data Institut Teknologi Sumatera yang berkontribusi dalam membantu dalam penulisan dan percobaan penelitian "Implementasi *Sentence Transformer* Berbasis BERT untuk Edukasi dan Informasi tentang HIV/AIDS" baik dukungan mental, ilmu dan materi, sehingga model dan penulisan penelitian ini dapat diselesaikan dengan tepat waktu dan mendapatkan hasil yang dapat dipertanggungjawabkan.

DAFTAR PUSTAKA

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/arXiv.1810.04805

Nastase, V., & Merlo, P. (2023). *Grammatical information in BERT sentence embeddings as two-dimensional arrays*. In Proceedings of the Workshop on Representation Learning for NLP (pp. 22-39). https://doi.org/10.18653/v1/2023.repl4nlp-1.3

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. https://doi.org/10.48550/arXiv.1908.10084

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT*. *Proceedings of the International Conference on Learning Representations* (ICLR). https://doi.org/10.48550/arXiv.1904.09675

L. Tunstall, N. Reimers, U. E. S. Jo andothers, "*Efficient Few-Shot Learning Without Prompts*," 2022. DOI: 10 . 48550 / ARXIV . 2209 . 11055. https://arxiv.org/abs/2209.11055.

Purnamawati, D., Supradewi, I., Sukamti, S., & Damayani, A. D. (2016). Pendidikan kesehatan: HIV dan AIDS bagi tenaga kesehatan. STIKes Kharisma Karawang. ISBN 978-602-60312- 1-1.

Firmanto, B., Aziz, A. S., & Sesoca, J. (2024, September 2). Tinjauan perkembangan kecerdasan buatan berbasis arsitektur Transformer. SMARTICS Journal, 10(1), 33–38.

Maulana, M. F. Z., Hasibuan, A., & Mauliah, S. (2024, Februari 1). Melonjaknya kasus HIV di kalangan remaja Indonesia. AMSIR *Community Service Journal*, 2(1), 1–7.

Purba, M. R., Errisya, M. K., Khofipah, S., & Purba, S. H. (2024). Analisis peran media sosial dalam meningkatkan kesadaran masyarakat tentang pencegahan penyakit HIV/AIDS. Jurnal Kesehatan Ilmiah Indonesia (*Indonesian Health Scientific Journal*), 9(1), 274–282.

JMLcommunity. (2020). Kurikulum tentang *Artificial Intelligence*. *Retrieved from* https://github.com/JMLcommunity/Kurikulum-AI

Kristiono, N., & Astuti, I. (2019). Mengenal HIV & AIDS. Universitas Negeri Semarang. Aljabar, A., & Karomah, B. M. (2024). Mengungkap opini publik: Pendekatan BERT-based-

caused untuk analisis sentimen pada komentar film. *Journal of System and Computer Engineering (JSCE)*, 5(1), 36.