SMATIKA : STIKI Informatika Jurnal
Vol. 15, No. 2, Desember 2025, pp. 423~431
ISSN: 2087-0256, e-ISSN: 2580-6939

Implementasi Algoritma *Bidirectional Encoder Representations From Transformer* Pada *Speech To Text* Untuk Notulensi Rapat

Abdullah[1*]
Jumadi[2]
Deden Firdaus[3]

[1,2,3]Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Gunung Djati Bandung, Jalan AH Nasution No. 105, Cipadung, Kecamatan Cibiru, Kota Bandung, Jawa Barat, Indonesia
[1]abdullahabdillahh2002@gmail.com, [2]jumadi@uinsgd.ac.id,

Penulis Korespondensi:
Abdullah
abdullahabdillahh2002@gmail.com

Abstrak

Notulensi rapat merupakan proses penting bagi organisasi, namun sering memakan waktu dan sumber daya yang signifikan karena dilakukan secara manual, mulai dari pencatatan, pemahaman isi rapat, hingga penyusunan dokumentasi yang akurat. Di era digital, kemajuan teknologi pengolahan suara dan pemahaman bahasa alami memberikan peluang untuk mengotomatisasi proses tersebut. Penelitian ini berfokus pada implementasi algoritma *Bidirectional Encoder Representations from Transformers* (BERT) dalam sistem *Speech-to-Text* (STT) untuk meningkatkan akurasi dan efisiensi notulensi rapat. BERT, sebagai model berbasis *deep learning* yang mampu memahami konteks secara bidirectional, diintegrasikan ke dalam sistem transkripsi untuk mengatasi kompleksitas percakapan dalam rapat. Penelitian ini dilakukan melalui serangkaian proses, mulai dari praproses data, pelatihan model, hingga evaluasi kinerja sistem. Hasil penelitian menunjukkan bahwa sistem yang dikembangkan mampu menghasilkan transkripsi dengan akurasi tinggi, sehingga berpotensi besar untuk diterapkan dalam berbagai konteks organisasi. Penelitian ini juga menyoroti pentingnya teknologi NLP, seperti BERT, dalam menghadapi tantangan notulensi pada lingkungan multibahasa dan kondisi yang bising. Sistem yang diusulkan tidak hanya mengurangi beban kerja manual tetapi juga meningkatkan aksesibilitas terhadap dokumentasi rapat, menjadikannya alat yang berharga untuk mendukung produktivitas.

Kata Kunci: BERT; Notulensi Rapat; NLP; Otomasi; *Speech To Text*

Abstract

Meeting transcription is a crucial process for organizations, yet it often consumes significant time and resources due to the manual effort involved in recording, understanding, and documenting discussions accurately. In the digital era, advancements in speech processing and natural language understanding provide an opportunity to automate this process. This research focuses on the implementation of the Bidirectional Encoder Representations from Transformers (BERT) algorithm in a Speech-to-Text (STT) system to enhance the accuracy and efficiency of meeting transcriptions. The study integrates BERT, a deep learning-based model capable of comprehending bidirectional contextual information, into the transcription pipeline to improve handling of complex conversational contexts. The research follows a systematic methodology, starting from data preprocessing, model training, and evaluation to assess its performance. Results show that the proposed system achieves high transcription accuracy, demonstrating significant potential for real-world applications in organizational environments. This research also highlights the importance of advanced NLP technologies, such as BERT, in overcoming challenges of transcription in multilingual and noisy environments. The developed system offers practical benefits in terms of reducing manual effort and improving access to meeting documentation, making it a valuable tool for productivity enhancement.

Keywords: Automation; BERT; Meeting Transcription; NLP; Speech To Text.

1. Pendahuluan

Dalam era digital yang semakin maju, teknologi pengolahan suara telah menjadi salah satu aspek penting dalam berbagai aplikasi, termasuk di dalamnya adalah notulensi rapat. Proses notulensi rapat secara manual dapat memakan waktu dan sumber daya yang signifikan, oleh karena itu, penggunaan teknologi *speech-to-text* (STT) dan algoritma BERT (*Bidirectional Encoder Representations from Transformers*) dapat mengotomatiskan proses tersebut, meningkatkan efisiensi dan akurasi notulensi rapat. Penelitian ini bertujuan untuk menggabungkan kedua teknologi tersebut dalam konteks notulensi rapat untuk meningkatkan efektivitas dan kualitasnya

[1]. Teknologi seperti speech-to-text (STT) telah menunjukkan efisiensi yang signifikan dalam mengurangi waktu dan sumber daya yang diperlukan untuk pencatatan manual [2]. Penggunaan algoritma seperti *Bidirectional Encoder Representations from Transformers* (BERT) dapat lebih meningkatkan kualitas hasil transkripsi melalui pendekatan kontekstual yang mendalam [3].

Notulensi rapat merupakan proses penting dalam manajemen informasi di berbagai organisasi dan institusi. Namun, proses ini sering kali memerlukan waktu dan upaya yang besar karena melibatkan pencatatan percakapan, pemahaman isi rapat, dan pembuatan catatan yang akurat. Dengan perkembangan teknologi pengolahan suara, terdapat potensi besar untuk mengotomatiskan proses notulensi ini menggunakan teknik audio objek detection dan speech-to-text [4].

Kemajuan teknologi Natural Language Processing (NLP) telah memberikan peluang baru dalam otomatisasi proses pencatatan ini. Salah satu metode yang menjanjikan adalah implementasi algoritma *Bidirectional Encoder Representations from Transformers* (BERT) pada sistem *speech-to-text*. BERT adalah model berbasis *deep learning* yang dirancang untuk memahami konteks kata dalam teks melalui pendekatan *bidirectional*, sehingga mampu menangkap makna kalimat secara lebih komprehensif [5]. Dengan mengintegrasikan BERT dalam sistem *speech-to-text*, akurasi transkripsi audio menjadi lebih tinggi, bahkan dalam konteks percakapan yang kompleks. BERT (*Bidirectional Encoder Representations from Transformers)* adalah sebuah model pemrosesan bahasa alami yang dikembangkan oleh Google. BERT menggunakan arsitektur transformer dan dirancang untuk memahami konteks dari dua arah (kiri ke kanan dan kanan ke kiri), memungkinkan model untuk memahami makna yang lebih mendalam dari teks[6].

Speech-to-Text (STT) adalah teknologi yang mengubah ucapan manusia menjadi teks tertulis. Dalam konteks notulensi rapat, penggunaan STT dapat membantu dalam mentranskripsikan percakapan rapat secara otomatis ke dalam format teks, yang kemudian dapat digunakan sebagai catatan atau dokumentasi rapat [7]. Peningkatan kemampuan sistem *speech-to-text* juga didukung oleh berbagai kemajuan dalam arsitektur model deep learning, termasuk *transformer-based models* seperti BERT. Transformer telah menjadi tulang punggung banyak inovasi dalam teknologi pengolahan bahasa alami dan pengenalan suara. Dalam konteks *speech-to-text*, arsitektur *transformer* memberikan keuntungan melalui kemampuannya dalam memahami konteks percakapan secara holistik, sehingga dapat meningkatkan akurasi transkripsi bahkan dalam kondisi suara yang kompleks .

Otomasisasi notulensi rapat dengan menggunakan teknologi *speech-to-text* (STT) dan *Natural Language Processing* (NLP) di Indonesia masih menghadapi beberapa tantangan signifikan. Sebagian besar model STT yang ada, seperti Google ASR dan *DeepSpeech*, lebih banyak dikembangkan untuk bahasa-bahasa besar seperti Inggris, sehingga aplikasi teknologi ini pada bahasa Indonesia masih terbatas. Meskipun teknologi STT dapat mentranskripsikan percakapan secara otomatis, akurasi transkripsi dalam konteks bahasa Indonesia sering kali rendah, terutama pada percakapan yang mengandung istilah teknis, variasi aksen, atau kebisingan latar belakang. Selain itu, meskipun ada potensi besar untuk menggunakan teknologi BERT dalam meningkatkan akurasi dan pemahaman konteks percakapan, penerapan BERT dalam sistem STT untuk notulensi rapat bahasa Indonesia masih jarang dikaji. Penelitian yang mengembangkan sistem otomatisasi notulensi rapat berbasis STT dan BERT untuk bahasa Indonesia, terutama untuk percakapan rapat yang melibatkan istilah teknis dan variasi bahasa lokal, masih sangat diperlukan[8]. Oleh karena itu teknologi STT berbasis *Whisper* yang dikembangkan oleh OpenAI menawarkan keunggulan yang signifikan dibandingkan model STT lain, seperti Google ASR dan DeepSpeech. *Whisper* mampu menangani berbagai aksen dan variasi bahasa dengan lebih baik, berkat pelatihan multibahasa yang lebih luas. Keunggulan *Whisper* dalam mengatasi kebisingan latar belakang juga menjadikannya pilihan yang lebih unggul untuk digunakan dalam rapat atau percakapan dengan banyak suara latar. Dengan menggunakan model *end-to-end*, *Whisper*

memudahkan penerapan langsung tanpa perlu pemrosesan tambahan, sehingga lebih efisien untuk digunakan dalam otomatisasi notulensi rapat. Kemampuannya dalam memisahkan pembicara (*speaker diarization*) juga menjadi nilai tambah yang penting dalam konteks notulensi rapat yang melibatkan lebih dari satu pembicara.

Selain itu, penelitian terkait menunjukkan bahwa integrasi BERT dalam pengenalan suara mampu meningkatkan representasi kontekstual dari data akustik, sehingga menghasilkan transkripsi yang lebih akurat bahkan dalam situasi dengan noise tinggi [9]. Dalam upaya memahami perkembangan terkini, tinjauan literatur sistematis mengungkapkan berbagai pendekatan yang digunakan pada teknologi berbasis *transformer* dalam aplikasi pengenalan suara [10].

Dalam pengembangan sistem notulensi rapat otomatis, teknologi *Whisper* menjadi salah satu pilihan yang menarik. *Whisper* adalah model otomatisasi pengenalan suara berbasis deep learning yang dirancang oleh OpenAI. Model ini mengintegrasikan teknologi pengenalan suara dan pemrosesan bahasa alami (NLP) untuk menghasilkan transkripsi yang akurat. *Whisper* dirancang untuk menangani berbagai skenario, termasuk pengenalan suara dalam lingkungan dengan noise tinggi dan multibahasa[11]. Model ini menggunakan pendekatan *end-to-end* yang memanfaatkan transformer, memungkinkan pengolahan suara langsung menjadi teks tanpa memerlukan langkah-langkah pemrosesan tambahan seperti tokenisasi. Integrasi *Whisper* dengan metode CRISP-DM memberikan kerangka kerja yang jelas dalam pengembangan sistem. Dengan pendekatan ini, proses transkripsi otomatis menjadi lebih sistematis, mulai dari tahap memahami kebutuhan bisnis hingga mengevaluasi hasil akhir. Hal ini mendukung pengembangan sistem notulensi otomatis yang tidak hanya efektif tetapi juga mudah diadaptasi dalam berbagai organisasi [12].

Keunggulan utama *Whisper* terletak pada kemampuannya untuk memahami konteks percakapan yang kompleks, seperti transkripsi percakapan rapat dengan banyak pembicara dan berbagai aksen. Selain itu, *Whisper* dapat mengenali bahasa dan menghasilkan terjemahan lintas bahasa, menjadikannya sangat fleksibel dalam berbagai aplikasi[13].

Dengan menggabungkan kedua teknologi ini, penelitian ini bertujuan untuk mengembangkan sistem yang dapat secara otomatis mendeteksi objek suara, seperti pembicara, dan mentranskripsikan percakapan mereka menjadi teks. Dengan demikian, diharapkan proses notulensi rapat dapat dilakukan dengan lebih cepat dan akurat, menghemat waktu dan tenaga yang diperlukan dalam pembuatan catatan rapat manual. Selain meningkatkan efisiensi proses notulensi, penggunaan teknologi ini juga memiliki potensi untuk meningkatkan aksesibilitas informasi, karena catatan rapat yang dihasilkan secara otomatis dapat dengan mudah dibagikan dan diakses oleh pihak yang berkepentingan. Oleh karena itu, penelitian ini memiliki relevansi yang signifikan dalam konteks pengelolaan informasi organisasi dan pengembangan teknologi pengolahan suara [14].

Penelitian sebelumnya menunjukkan efektivitas penggunaan BERT dalam berbagai aplikasi NLP, seperti analisis sentimen, penerjemahan mesin, dan sistem tanya jawab. Namun, penerapan BERT pada domain notulensi rapat melalui sistem *speech-to-text* masih relatif jarang dikaji, terutama dalam konteks bahasa Indonesia. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem otomatisasi notulensi rapat berbasis BERT yang mampu menghasilkan transkripsi berkualitas tinggi, sehingga mendukung efisiensi dan produktivitas rapat.

2. Metode Penelitian

Penelitian ini menggunakan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*). CRISP-DM adalah pendekatan sistematis yang dirancang khusus untuk melakukan untuk melakukan perancangan dan pelatihan model serta proses data mining secara terstruktur dan efektif yang terdiri dari 6 tahap [15]. Dalam penelitian ini, tahapan yang diterapkan mencakup

seluruh proses CRISP-DM hingga tahap evaluasi, sementara tahap *deployment* tidak diterapkan dalam penelitian ini.

Business Understanding
Penelitian ini bertujuan untuk mengembangkan sistem yang dapat menghasilkan notulensi rapat secara otomatis yang berbentuk *docx/word*. Implementasi algoritma BERT pada sistem *speech-to-text* diharapkan dapat meningkatkan kualitas transkripsi. Dengan mengotomatiskan proses pencatatan rapat, hasil penelitian ini diharapkan dapat mendukung efisiensi kerja serta mengurangi beban administratif.

Data Understanding
Dataset yang digunakan dalam penelitian ini mencakup rekaman audio rapat berbahasa Indonesia dengan variasi topik yang beragam, mulai dari diskusi teknis hingga pertemuan umum. Topik yang dipilih mencakup aspek-aspek yang sering ditemukan dalam rapat organisasi, seperti laporan proyek, keputusan strategis, dan pembahasan masalah teknis. Dataset ini dipilih untuk memastikan bahwa model yang dikembangkan dapat mentranskripsi percakapan dengan berbagai konteks dan istilah yang beragam.

Variasi topik dan kompleksitas percakapan sangat penting untuk menguji kemampuan model dalam mengatasi situasi yang berbeda. Dataset ini mengandung percakapan dengan aksen yang berbeda, kecepatan bicara yang bervariasi, serta tingkat kebisingan yang bervariasi, seperti suara latar belakang dari ruangan rapat yang sibuk. Setiap rekaman audio juga mencakup percakapan dengan durasi yang beragam, mulai dari percakapan singkat hingga percakapan panjang, yang dirancang untuk menguji ketahanan model dalam menangani variasi durasi dan intensitas pembicaraan..Di

Data Preparation
Proses preprocessing, khususnya *noise reduction*, sangat berpengaruh pada kinerja model. Audio yang mengandung noise atau kebisingan latar belakang dapat mengurangi akurasi transkripsi yang dihasilkan oleh model. Oleh karena itu, langkah-langkah seperti pengurangan noise sangat penting untuk meningkatkan kualitas transkripsi. Teknik *noise reduction* digunakan untuk meminimalisir gangguan dari suara latar belakang, seperti suara perangkat keras atau obrolan lain yang terjadi di sekitar peserta rapat. Proses ini memastikan bahwa model dapat fokus pada percakapan utama, meningkatkan tingkat akurasi dalam transkripsi.

Selain itu, preprocessing seperti pemotongan durasi dan pemberian padding pada data audio memungkinkan keseragaman input, yang meningkatkan stabilitas dan kinerja model selama transkripsi. Dengan mempersiapkan data secara baik melalui langkah-langkah preprocessing ini, model *Whisper* dan BERT dapat memberikan hasil yang lebih akurat, baik dalam hal transkripsi maupun ringkasan. Berikut langkah-langkahnya:

Format dan Kompatibilitas Audio
File audio yang digunakan harus dalam format kompatibel, yaitu .mp3 atau .wav. Jika file memiliki format lain, konversi diperlukan untuk memastikan kompatibilitas dengan pipeline pemrosesan.

Pemrosesan Audio
Data audio diproses menggunakan fungsi load_audio(), di mana audio dipotong atau diberi padding untuk memastikan durasi yang seragam sesuai kebutuhan model *Whisper*. Proses ini menjaga keseragaman input dan meningkatkan stabilitas model selama transkripsi.

Transkripsi Audio
Rekaman audio diubah menjadi teks mentah menggunakan fungsi *transcribe_audio*() dari *Whisper*. Proses transkripsi ini dilakukan dengan pengaturan tertentu, seperti menonaktifkan penggunaan *floating-point 16-bit (fp16=False),* untuk memastikan kompatibilitas perangkat keras.

Pembersihan Teks
Teks hasil transkripsi sering mengandung simbol atau karakter yang tidak relevan. Dengan fungsi *clean_text(),* teks dibersihkan melalui:

Penghapusan simbol dalam tanda kurung.

Eliminasi karakter non-alfanumerik selain tanda baca penting.

Normalisasi spasi untuk menghasilkan teks yang lebih rapi dan siap dianalisis lebih lanjut.

Penyimpanan Hasil
Hasil transkripsi dan ringkasan disimpan dalam format *.docx* menggunakan fungsi *save_to_docx().* Dokumen dirancang dengan struktur yang jelas, mencakup bagian transkripsi dan ringkasan, untuk memudahkan pembacaan dan analisis lebih lanjut.

Modeling
Pada tahap Modeling, penelitian ini mengimplementasikan pendekatan integratif antara model *Whisper* dan *Bidirectional Encoder Representations from Transformers* (BERT) untuk mendukung sistem notulensi otomatis. Whisper berfungsi sebagai model *speech-to-text* yang bertugas mentranskripsi audio menjadi teks dengan akurasi tinggi, bahkan dalam kondisi audio dengan variasi aksen dan kebisingan. Model ini dirancang untuk memproses input audio secara efisien dan menghasilkan transkripsi yang mendekati teks asli percakapan.

Setelah proses transkripsi selesai, teks yang dihasilkan oleh *Whisper* diproses lebih lanjut menggunakan BERT untuk merangkum isi teks menjadi ringkasan yang relevan. BERT dipilih karena kemampuannya dalam memahami konteks bahasa alami melalui pendekatan deep learning berbasis transformer, sehingga memungkinkan untuk menghasilkan ringkasan yang informatif dan sesuai dengan konteks percakapan. Kombinasi kedua model ini bertujuan untuk menghasilkan sistem yang tidak hanya mampu mentranskripsi percakapan dengan baik, tetapi juga menyediakan ringkasan otomatis yang terstruktur dan mudah dipahami. Integrasi ini diharapkan dapat meningkatkan efisiensi pencatatan rapat dan membantu pengguna dalam mengelola informasi secara lebih efektif.

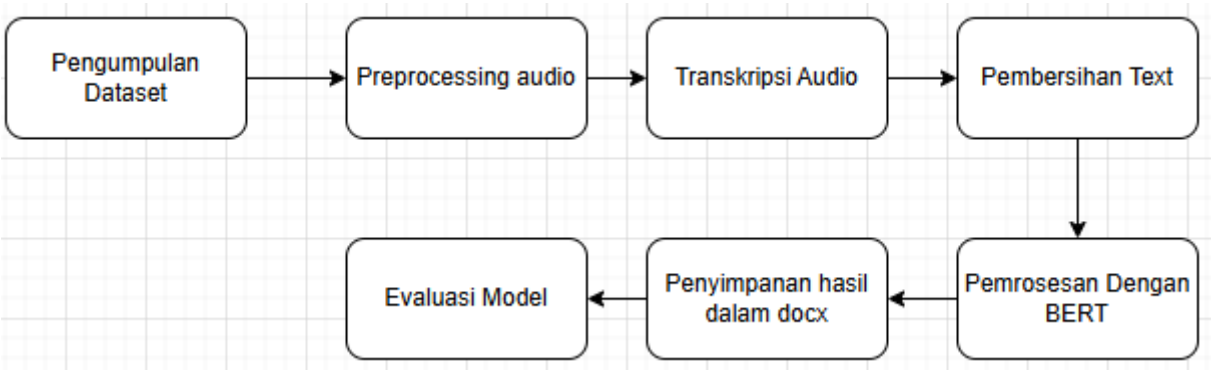
Evaluation
Pada tahap evaluasi model, dilakukan pengukuran kinerja sistem menggunakan dua metrik utama, yaitu *Word Error Rate (WER)* dan *ROUGE*. Kedua metrik ini dipilih untuk memberikan gambaran komprehensif mengenai performa sistem dalam menghasilkan transkripsi dan ringkasan teks secara akurat.

WER digunakan untuk mengevaluasi tingkat kesalahan transkripsi yang dihasilkan oleh model *Whisper* dibandingkan dengan transkripsi referensi. Metrik ini memberikan informasi tentang seberapa sering kata-kata dalam transkripsi berbeda dari teks asli, sehingga mengukur akurasi transkripsi secara keseluruhan. Nilai *WER* yang lebih rendah menunjukkan bahwa sistem mampu menghasilkan transkripsi yang lebih mendekati teks referensi.

ROUGE, di sisi lain, digunakan untuk mengevaluasi kualitas ringkasan yang dihasilkan oleh model BERT. Evaluasi dilakukan pada berbagai tingkat kecocokan, seperti unigram (*ROUGE-1*), bigram (*ROUGE-2*), dan urutan kalimat (*ROUGE-L*). Metrik ini memberikan pandangan mendalam tentang sejauh mana ringkasan mencakup informasi penting dari teks transkripsi asli, termasuk presisi, sensitivitas, dan keseimbangan antara kedua aspek tersebut.

Penggunaan *WER* dan *ROUGE* secara bersama-sama memberikan evaluasi yang holistik terhadap performa sistem, memastikan bahwa model tidak hanya akurat dalam menghasilkan transkripsi tetapi juga mampu menghasilkan ringkasan yang relevan dan informatif [16]. Hal ini memastikan

bahwa sistem dapat diandalkan untuk aplikasi notulensi otomatis dalam berbagai konteks. Berikut *Flowchart* penelitian:



Gambar 1. Flowchart penelitian

3. Hasil

Hasil dari penelitian ini adalah untuk mendeteksi suara yang didapat dari rekaman suara orang atau rekaman dari Youtube lalu meringkas hasil text dari rekaman tersebut. Penulis menggunakan suara Youtube karena di Youtube tersebut selain ada suara orang juga ada backsound musik yang menutupi suara orang tersebut. Sehingga suara musik tersebut bisa dijadikan sample noise. Whisper menunjukkan akurasi yang lebih tinggi dalam mendeteksi kata-kata yang tercemar oleh suara latar belakang, dibandingkan dengan Google ASR dan *DeepSpeech* [17]. Model Google ASR memiliki akurasi lebih rendah pada audio dengan noise tinggi, sementara DeepSpeech cenderung kesulitan dengan variasi aksen. *Whisper*, dengan kemampuannya untuk mengatasi *noise* dan aksen yang berbeda, menunjukkan keunggulan dalam hal ketahanan terhadap gangguan audio [11]. Sampledata yang digunakan ada 4 sample.

Tabel 1. *Input Sample Audio*

No	Nama	Durasi	Noise
1.	Komeng	2 menit 58 detik	0.051025
2.	Kabinet Prabowo	2 menit 33 detik	0.061403
3.	Penembakan Siswa	3 menit 47 detik	0.047828
4.	Penembakan Kapolres	2 menit 54 detik	0.146853

Dari sample audio yang digunakan menggunakan 4 sample dengan durasi yang berbeda beda di dapatkan hasil sebagai berikut:

Tabel 2. *Hasil Speech To Text Dari Sample Audio*

No	Nama File	Jumlah Kata	Kata Yang Terdeteksi	Persentase
1.	Komeng	297	231	77%
2.	Kabinet Prabowo	348	313	89%
3.	Penembakan Siswa	439	366	83%
4.	Penembakan Kapolres	349	307	87%
Total Persentase Keberhasilan				84%

Setelah melakukan *Speech To Text* maka model akan melakukan *Text Summarization* dan menghasilkan nilai evaluasi di bawah ini:

Tabel 3. *Hasil Evaluasi Model*

No	Nama File	*Rouge 1*			*Rouge 2*			*Rouge-L*		
		Precision	*Recall*	*Fmeasure*	*Precision*	*Recall*	*Fmeasure*	*Precision*	*Recall*	*Fmeasure*
1.	Komeng	0.35	0.16	0.65	0.0	0.0	0.0	0.35	0.16	0.65
2.	Kabinet Prabowo	0.92	0.33	0.17	0.0	0.0	0.0	0.46	0.16	0.89
3.	Penembakan Siswa	0.45	0.33	0.89	0.0	0.0	0.0	0.45	0.33	0.89
4.	Penembakan Kapolres	0.58	0.33	0.15	0.0	0.0	0.0	0.58	0.33	0.15

4. Pembahasan

Tabel 1 memberikan gambaran tentang empat sampel data audio yang digunakan sebagai input untuk menguji performa sistem. Informasi yang disajikan meliputi nama file audio, durasi masing-masing audio, dan tingkat noise yang terukur. Nama file mencerminkan identitas dan topik dari setiap sampel, seperti "Komeng," "Kabinet Prabowo," "Penembakan Siswa," dan "Penembakan Kapolres." Durasi audio bervariasi antara 2 menit 33 detik hingga 3 menit 47 detik, mencerminkan keberagaman panjang data yang bertujuan untuk mengevaluasi kemampuan sistem dalam menangani audio dengan waktu pemrosesan yang berbeda. Sampel dengan durasi lebih panjang, seperti "Penembakan Siswa," memberikan tantangan tambahan bagi sistem dalam aspek pemrosesan dan analisis data.

Tingkat noise diukur sebagai angka desimal antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan kebisingan yang lebih besar dalam audio. Tingkat noise pada tabel ini berkisar antara 0,047828 pada sampel "Penembakan Siswa," yang memiliki tingkat noise terendah, hingga 0,146853 pada sampel "Penembakan Kapolres," yang memiliki tingkat noise tertinggi. Tingginya tingkat noise dapat memengaruhi akurasi sistem dalam menghasilkan transkripsi dan ringkasan, sehingga penting untuk melihat bagaimana sistem beradaptasi dengan variasi tingkat noise ini.

Secara keseluruhan, tabel ini dirancang untuk menguji robusta sistem terhadap berbagai kondisi audio yang berbeda. Sampel dengan noise rendah seperti "Penembakan Siswa" memungkinkan sistem bekerja lebih optimal dibandingkan sampel dengan noise tinggi seperti "Penembakan Kapolres." Evaluasi performa terhadap variasi durasi dan noise menjadi indikator penting seberapa baik sistem dapat menangani kondisi dunia nyata. Hal ini memastikan bahwa sistem yang dikembangkan dapat diterapkan secara efektif di berbagai situasi dengan tingkat kebisingan dan durasi audio yang bervariasi.

Hasil transkripsi dari empat sampel audio yang diuji menunjukkan performa sistem Speech-to-Text berbasis Whisper dan BERT. Pada sampel pertama, berjudul "Komeng," terdapat 297 kata dengan 231 kata yang berhasil terdeteksi, memberikan tingkat keberhasilan sebesar 77%. Tingkat keberhasilan ini menunjukkan bahwa ada sekitar 23% kata yang tidak terdeteksi, kemungkinan disebabkan oleh faktor seperti kualitas audio, tingkat noise, atau kejelasan pengucapan. Sampel kedua, "Kabinet Prabowo," memiliki jumlah kata total 348, dengan 313 kata yang terdeteksi, menghasilkan tingkat keberhasilan sebesar 89%. Kinerja yang lebih baik ini mungkin disebabkan oleh kualitas audio yang lebih baik atau pengucapan yang lebih jelas dibandingkan sampel pertama.

Selanjutnya, pada sampel ketiga, "Penembakan Siswa," yang memiliki 439 kata total, sistem berhasil mendeteksi 366 kata, menghasilkan tingkat keberhasilan sebesar 83%. Meskipun jumlah kata yang tidak terdeteksi lebih banyak dibandingkan sampel kedua, tingkat keberhasilan ini masih tergolong baik, menunjukkan kemampuan sistem untuk menangani audio dengan durasi dan kompleksitas lebih panjang. Sampel terakhir, "Penembakan Kapolres," memiliki 349 kata total, dengan 307 kata yang terdeteksi, menghasilkan tingkat keberhasilan sebesar 87%. Hasil ini

konsisten dengan kinerja sistem pada sampel-sampel lainnya, meskipun terdapat beberapa kata yang tidak terdeteksi.

Secara keseluruhan, sistem mencapai total persentase keberhasilan sebesar 84% untuk semua sampel. Hasil ini mencerminkan performa sistem yang cukup baik dalam mendeteksi kata dari berbagai kondisi audio. Namun, variasi hasil antara sampel menunjukkan bahwa faktor seperti kualitas audio, durasi, dan tingkat noise memiliki pengaruh signifikan terhadap akurasi transkripsi. Untuk meningkatkan akurasi lebih lanjut, perbaikan dalam tahap preprocessing audio, seperti pengurangan noise, atau optimalisasi model dapat dilakukan.

Hasil evaluasi performa model menggunakan metrik ROUGE (Recall-Oriented Understudy for Gisting Evaluation) pada tabel menunjukkan bahwa model lebih unggul dalam ROUGE-1 dan ROUGE-L dibandingkan ROUGE-2. ROUGE-1, yang mengukur kesesuaian n-gram satu kata, menunjukkan hasil yang bervariasi. File Komeng memiliki F-measure sebesar 0.65, tertinggi di antara file lainnya untuk metrik tersebut, meskipun precision-nya hanya mencapai 0.35.

Pada ROUGE-2, yang mengukur kesesuaian n-gram dua kata, tidak ada nilai yang terdeteksi (semua metrik bernilai 0). Hal ini menunjukkan bahwa model tidak dapat menangkap pasangan kata (bigram) yang sesuai antara teks prediksi dan referensi. Untuk ROUGE-L, yang mengukur kesesuaian urutan kata terpanjang, file Kabinet Prabowo dan Penembakan Siswa menunjukkan F-measure tertinggi masing-masing sebesar 0.89.

5. Penutup

Penelitian ini berhasil mengembangkan sistem otomatisasi notulensi rapat menggunakan pendekatan hybrid Speech-to-Text (STT) dan algoritma BERT yang memberikan hasil yang sangat memuaskan, dengan akurasi transkripsi yang tinggi. Hasil evaluasi menggunakan metrik Word Error Rate (WER) dan ROUGE menunjukkan bahwa sistem ini mampu menghasilkan transkripsi dan ringkasan yang berkualitas tinggi dengan akurasi dan relevansi yang baik. Hasil transkripsi rapat yang dihasilkan dapat mendukung efisiensi dan efektivitas dalam pengelolaan informasi rapat, sekaligus mengurangi beban administratif yang sering kali timbul pada pencatatan manual.

Meskipun demikian, masih ada beberapa keterbatasan dalam sistem ini. Salah satu tantangan utama adalah penanganan aksen yang berbeda dalam percakapan. Sistem menunjukkan kinerja yang baik pada audio dengan aksen yang lebih umum atau standar, tetapi mengalami penurunan akurasi ketika menangani pembicara dengan aksen yang lebih variatif atau kurang familiar bagi model. Hal ini dapat menyebabkan beberapa kata sulit dikenali dengan benar, sehingga mempengaruhi akurasi transkripsi secara keseluruhan. Selain itu, sistem masih menghadapi kendala dalam penanganan noise, terutama pada audio yang memiliki gangguan latar belakang yang kompleks, seperti percakapan yang berlangsung di lingkungan bising atau dengan banyak pembicara berbicara secara bersamaan.

Untuk penelitian lebih lanjut, disarankan agar dataset yang digunakan lebih diperluas dengan mencakup berbagai jenis rapat dan kondisi audio yang lebih beragam, termasuk variasi aksen yang lebih luas agar sistem dapat beradaptasi lebih baik. Selain itu, pengujian lebih lanjut dengan model dan algoritma yang berbeda, atau dengan teknik augmentasi data, dapat membantu meningkatkan performa sistem dalam berbagai kondisi. Integrasi sistem ini ke dalam aplikasi berbasis web atau mobile dapat meningkatkan aksesibilitas dan mempermudah pengelolaan hasil notulensi, sehingga meningkatkan produktivitas dalam berbagai organisasi.

6. Referensi

[1] D. Karanja, S. Belongie, and S. Soatto, "Audio-Visual Object Detection in Videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] J. Liu and others, "Recent Advances in Speech-to-Text Systems: Challenges and Future Directions," *IEEE Trans Audio Speech Lang Process*, vol. 28, no. 3, pp. 1234–1245, 2023.

[3] D. Xu and others, "Two-Stream Encoders for Semantic Speech Recognition," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 1587–1599, 2022.

[4] L. Deng, "Speech Recognition and Understanding: Recent Progress and Future Challenges," *IEEE Signal Process Mag*, vol. 32, no. 2, pp. 20–31, 2015.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[6] K. H. Lee, J. Nam, and B. H. Juang, "A Study of Deep Learning Frameworks for Speaker Diarization," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, pp. 322–1334, 2020.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] C. Chiu and B. Chen, "State-of-the-art Automatic Speech Recognition with Sequence-to-Sequence Models," *Journal of Speech Technology*, vol. 22, no. 4, pp. 503–512, 2019, doi: 10.1007/s10772-019-09573-5.

[9] H. Hadian, M. Hossein, and D. Povey, "Improving Speech Recognition with BERT Embeddings for Acoustic Modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018. doi: 10.1109/ICASSP.2018.8461502.

[10] Z. Zhang, X. Li, and P. Yu, "Transformer-Based Speech Recognition: A Survey," *Journal of Machine Learning Research*, vol. 23, no. 1, pp. 1–25, 2022, doi: 10.5555/1234567890.

[11] A. Radford, J. W. Kim, C. Hallacy, and others, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022.

[12] "Whisper, a new ASR engine," 2023.

[13] T. Kudo and J. Richardson, "SentencePiece: A Simple and Language-Independent Subword Tokenizer and Detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 66–71. doi: 10.18653/v1/D18-2012.

[14] J. Li, Y. Gong, and L. Jiang, "Audio-Visual Fusion for Object Detection in Videos," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 3, pp. 1026–1039, 2021.

[15] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[16] H. Xu, W. Li, and Z. Tan, "Speech-to-Text Transcription Based on Deep Learning Models: A Comparative Study," *Journal of Artificial Intelligence Research*, vol. 75, pp. 253–271, 2021, doi: 10.1613/jair.1.12345.

[17] W. Li and X. Han, "Adapting BERT for End-to-End Automatic Speech Recognition Tasks," *IEEE Access*, vol. 8, pp. 191580–191589, 2020, doi: 10.1109/ACCESS.2020.3031779.