

**PENCOCOKAN JUDUL DAN TRANSKRIP VIDEO MENGGUNAKAN
METODE *COSINE SIMILARITY***

SKRIPSI

Oleh :
SYAFRI MAULANA ALFATIH
NIM. 220605110139



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

**PENCOCOKAN JUDUL DAN TRANSKRIP VIDEO MENGGUNAKAN
METODE *COSINE SIMILARITY***

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
SYAFRI MAULANA ALFATIH
NIM. 220605110139

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026**

HALAMAN PERSETUJUAN

**PENCOCOKAN JUDUL DAN TRANSKRIP VIDEO MENGGUNAKAN
METODE *COSINE SIMILARITY***

SKRIPSI

Oleh :
SYAFRI MAULANA ALFATHI
NIM. 220605110139

Telah Diperiksa dan Disetujui untuk Diuji:

Tanggal: 10 Juni 2026

Pembimbing I,



Dr. Ir. Fachrul Kurniawan, ST., M.MT., IPU
NIP. 19771020 200912 1 001

Pembimbing II,



Fatchurrochman, M.Kom
NIP. 19700731 200501 1 002

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

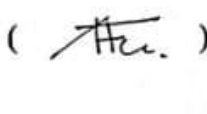
PENCOCOKAN JUDUL DAN TRANSKRIP VIDEO MENGGUNAKAN METODE *COSINE SIMILARITY*

SKRIPSI

Oleh :
SYAFRI MAULANA ALFATHI
NIM. 220605110139

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 12 Juni 2026

Susunan Dewan Penguji

Ketua Penguji	: <u>Hani Nurhayati, M.T</u> NIP. 19780625 200801 2 006	()
Anggota Penguji I	: <u>Tri Mukti Lestari, M.Kom</u> NIP. 19911108 202012 2 005	()
Anggota Penguji II	: <u>Dr. Ir. Fachrul Kurniawan, ST., M.MT., IPU</u> NIP. 19771020 200912 1 001	()
Anggota Penguji III	: <u>Fatchurrochman, M.Kom</u> NIP. 19700731 200501 1 002	()

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M. Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Syafri Maulana Alfatih
NIM : 220605110139
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Pencocokan Judul dan Transkrip Video
Menggunakan Metode Cosine Similarity

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 18 Juni 2026
Yang membuat pernyataan,



Syafri Maulana Alfatih
NIM. 220605110139

MOTTO

Ngomong aja dulu,

Mimpi aja dulu,

Bangunnya nanti,

Kalau nyawa sudah kumpul,

Dan badan siap tempur.

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur, karya ini penulis persembahkan kepada:

Allah SWT

Atas segala rahmat, nikmat, kesehatan, kemudahan, dan petunjuk-Nya sehingga penulis mampu menyelesaikan perjalanan ini dengan penuh rasa syukur.

Ayah, Mama, dan adik-adik tercinta

Atas doa, kasih sayang, dukungan, pengorbanan, serta kepercayaan yang selalu diberikan kepada penulis dalam setiap langkah dan perjuangan hingga terselesaikannya skripsi ini.

Seluruh keluarga besar

Yang senantiasa memberikan doa, dukungan, dan motivasi sehingga penulis dapat terus berusaha dan berkembang.

Dosen pembimbing, penguji, dan para pengajar

Atas bimbingan, arahan, ilmu, serta pengalaman berharga yang telah diberikan selama masa perkuliahan dan penyusunan skripsi ini.

Teman-teman Infinity Teknik Informatika 2022

Yang telah menjadi bagian dari perjalanan perkuliahan ini melalui kebersamaan,
dukungan, dan semangat yang tidak pernah padam.

Sahabat dan semua pihak yang telah membantu

Yang dengan tulus memberikan bantuan, dukungan, serta motivasi selama proses
penyusunan skripsi ini.

Terakhir, untuk diri sendiri yang sering terlupakan

Terima kasih telah berusaha, bertahan, dan tidak menyerah dalam menghadapi
berbagai tantangan hingga mampu menyelesaikan tanggung jawab ini dengan
baik.

KATA PENGANTAR

Assalamu 'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahirabbil'alam, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan rahmat, hidayah, kesehatan, dan kemudahan-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Pencocokan Judul dan Transkrip Video Menggunakan Metode Cosine Similarity" dengan baik. Shalawat serta salam senantiasa tercurah kepada junjungan Nabi Besar Muhammad SAW yang telah membawa umat manusia dari zaman kegelapan menuju zaman yang penuh dengan ilmu pengetahuan dan peradaban.

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Dalam proses penyusunan skripsi ini, penulis menyadari bahwa banyak pihak yang telah memberikan bantuan, dukungan, bimbingan, serta doa sehingga skripsi ini dapat terselesaikan dengan baik. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., selaku Rektor Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
2. Dr. H. Agus Mulyono, S.Pd., M.Kes, selaku Dekan Fakultas Sains dan Teknologi UIN Maulana Malik Ibrahim Malang.
3. Supriyono, M. Kom, selaku Ketua Program Studi Teknik Informatika UIN Maulana Malik Ibrahim Malang.

4. Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU, selaku Dosen Pembimbing I, yang dengan kesabaran dan dedikasinya membimbing penulis selama penyusunan skripsi.
5. Fatchurrochman, M.Kom., selaku Dosen Pembimbing II, yang senantiasa memberikan arahan dan masukan yang sangat berharga dalam pengerjaan skripsi ini.
6. Hani Nurhayati, M.T., dan Tri Mukti Lestari, M.Kom., selaku Dosen Penguji yang telah memberikan kritik, saran, dan bimbingan untuk menyempurnakan skripsi ini.
7. Nia Faricha, S.Si., admin Program Studi Teknik Informatika, yang membantu dalam urusan administrasi dan selalu mengingatkan tentang kelengkapan berkas.
8. Ayah, Mama, dan adik-adik tercinta yang selalu memberikan doa, kasih sayang, dukungan, semangat, serta motivasi yang tiada henti kepada penulis.
9. Seluruh keluarga besar yang senantiasa memberikan doa dan dukungan selama proses perkuliahan hingga penyelesaian skripsi ini.
10. Teman solid “PDI” yang beranggotakan Mumtaz, Sada, Maul, dan Yudha. Yang kebersamaan baik ketika berada dikala senang maupun sedih.
11. Teman-teman Infinity Teknik Informatika 2022 yang telah menjadi bagian dari perjalanan perkuliahan penulis melalui kebersamaan, kerja sama, serta dukungan yang tidak terlupakan.
12. Seluruh sahabat, rekan, dan pihak-pihak lain yang tidak dapat disebutkan

satu per satu yang telah membantu dan memberikan dukungan dalam penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki berbagai keterbatasan dan belum sepenuhnya sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan sebagai bahan evaluasi dan penyempurnaan di masa mendatang. Terlepas dari segala kekurangan yang ada, proses penyusunan skripsi ini telah menjadi pengalaman berharga yang memberikan banyak pembelajaran, baik secara akademis maupun pribadi. Penulis berharap penelitian ini dapat memberikan manfaat bagi pembaca, menjadi referensi bagi penelitian selanjutnya, serta memberikan kontribusi dalam pengembangan ilmu pengetahuan, khususnya pada bidang pengolahan bahasa alami dan analisis kesamaan teks. Semoga segala upaya yang telah dilakukan dalam penyusunan skripsi ini dapat bernilai kebaikan dan menjadi langkah awal untuk terus belajar, berkembang, serta berkarya di masa yang akan datang.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Malang, 18 Juni 2026

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN.....	iv
PERNYATAAN KEASLIAN TULISAN	v
MOTTO	vi
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	ix
DAFTAR ISI.....	xii
DAFTAR GAMBAR.....	xiv
DAFTAR TABEL	xv
ABSTRAK	xvi
ABSTRACT	xvii
البحث مستخلص.....	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Pernyataan Masalah	4
1.3 Batasan Masalah.....	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
BAB II STUDI PUSTAKA	7
2.1 Penelitian Terdahulu	7
2.2 Transkrip Video	11
2.3 <i>Preprocessing</i> Teks	12
2.4 <i>Sentence BERT</i>	12
2.5 <i>Cosine Similarity</i>	14
2.6 Metrik Evaluasi Klasifikasi.....	16
2.6.1 Confusion Matrix	17
2.6.1 Accuracy	18
2.6.1 Precision.....	18
2.6.1 Recall	19
2.6.1 F1-Score.....	20
BAB III METODOLOGI PENELITIAN	21
3.1 Sumber Data	21
3.2 Labelisasi Data	23
3.3 Perancangan Sistem	25
3.4 <i>Preprocessing</i>	26
3.4.1 Text Cleaning	27
3.4.2 Case Folding	27
3.5 Perhitungan <i>Sentence BERT</i>	28
3.5.1 Tokenisasi Sentence BERT	29
3.5.2 Token Embedding.....	31

3.5.3 Mean Pooling	32
3.6 <i>Cosine Similarity</i>	34
3.7 Contoh Perhitungan Lengkap	38
3.8 Rancangan Pengujian	40
3.9 Tahap Evaluasi Model	42
BAB IV HASIL DAN PEMBAHASAN	43
4.1 Persiapan Dataset	43
4.2 Hasil Labelisasi Data	45
4.3 Hasil <i>Preprocessing</i>	47
4.3.1 Hasil Cleaning	47
4.3.2 Hasil Case Folding	51
4.4 Hasil dan Analisis Uji Coba	55
4.4.1 Hasil Perhitungan Sentence BERT	55
4.4.2 Hasil Perhitungan Cosine Similarity	62
4.4.3 Skenario Pengujian	64
4.5 Evaluasi Kinerja Sistem	84
4.5.1 Confusion Matrix	84
4.5.2 Hasil Evaluasi dan Perbandingan Kinerja	90
4.6 Analisis Hasil	93
4.7 Integrasi Islam	99
BAB V KESIMPULAN DAN SARAN	103
5.1 Kesimpulan	103
5.2 Saran	103
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 2. 1 Konsep Dasar Cosine Similarity	15
Gambar 3. 1 Perancangan Sistem.....	25
Gambar 3. 2 Diagram Preprocessing	26
Gambar 4. 1 Distribusi Lebalisasi Data	46
Gambar 4. 2 Tampilan Hasil Cleaning pada Aplikasi.....	51
Gambar 4. 3 Tampilan Hasil Case Folding pada Aplikasi.....	54
Gambar 4. 4 Tampilan Input pada Aplikasi	55
Gambar 4. 5 Pseudocode Tokenisasi	56
Gambar 4. 6 Tampilan Hasil Tokenisasi pada Aplikasi.....	57
Gambar 4. 7 Pseudocode Token Embedding	58
Gambar 4. 8 Tampilan Hasil Token Embedding pada Aplikasi	59
Gambar 4. 9 Pseudocode Mean Pooling	60
Gambar 4. 10 Tampilan Hasil Mean Pooling pada Aplikasi	61
Gambar 4. 11 Pseudocode Cosine Similarity.....	62
Gambar 4. 12 Tampilan Hasil Cosine Similarity pada Aplikasi.....	63
Gambar 4. 13 Tampilan Hasil Output pada Aplikasi.....	64
Gambar 4. 14 Pseudocode Pemilihan Threshold	65
Gambar 4. 15 Hasil Eksperimen Skenario Pengujian 0.10 pada Aplikasi	66
Gambar 4. 16 Hasil Eksperimen Skenario Pengujian 0.15 pada Aplikasi	70
Gambar 4. 17 Hasil Eksperimen Skenario Pengujian 0.20 pada Aplikasi	73
Gambar 4. 18 Hasil Eksperimen Skenario Pengujian 0.25 pada Aplikasi	76
Gambar 4. 19 Hasil Eksperimen Skenario Pengujian Threshold Mean.....	80
Gambar 4. 20 Confusion Matrix pada Threshold 0.10.....	85
Gambar 4. 21 Confusion Matrix pada Threshold 0.15.....	86
Gambar 4. 22 Confusion Matrix pada Threshold 0.20.....	87
Gambar 4. 23 Confusion Matrix pada Threshold 0.25.....	88
Gambar 4. 24 Confusion Matrix pada Threshold 0.34 (Mean).....	89
Gambar 4. 25 Grafik Perbandingan Kinerja Model	92
Gambar 4. 26 Distribusi Similarity Sentence BERT	98

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	8
Tabel 2. 2 Confusion Matrix	17
Tabel 3. 1 Contoh Data	22
Tabel 3. 2 Contoh Data Label	23
Tabel 3. 3 Hasil Cleaning Q1 dan D1	27
Tabel 3. 4 Hasil Cleaning Q2 dan D2	27
Tabel 3. 5 Hasil Case Folding Q1 dan D1	28
Tabel 3. 6 Hasil Case Folding Q2 dan D2	28
Tabel 3. 7 Tokenisasi Sentence BERT Q1 dan D1	30
Tabel 3. 8 Tokenisasi Sentence BERT Q2 dan D2	30
Tabel 3. 9 Token Embedding Sentence BERT pada Q1	31
Tabel 3. 10 Token Embedding Sentence BERT pada D1	31
Tabel 3. 11 Token Embedding Sentence BERT pada Q2	31
Tabel 3. 12 Token Embedding Sentence BERT pada D2	31
Tabel 3. 13 Mean Pooling Sentence BERT Q1 dan D1	33
Tabel 3. 14 Mean Pooling Sentence BERT Q2 dan D2	33
Tabel 3. 15 Skenario Pengujian dengan Threshold.....	41
Tabel 4. 1 Contoh Dataset.....	44
Tabel 4. 2 Hasil Cleaning Judul pada Contoh Dataset.....	48
Tabel 4. 3 Hasil Cleaning Transkrip pada Contoh Dataset.....	49
Tabel 4. 4 Hasil Case Folding Judul pada Contoh Dataset	52
Tabel 4. 5 Hasil Case Folding Transkrip pada Contoh Dataset	53
Tabel 4. 7 Hasil Klasifikasi Data Berdasarkan Threshold 0.10	67
Tabel 4. 8 Hasil Klasifikasi Data Berdasarkan Threshold 0.15	71
Tabel 4. 9 Hasil Klasifikasi Data Berdasarkan Threshold 0.20	74
Tabel 4. 10 Hasil Klasifikasi Data Berdasarkan Threshold 0.25	77
Tabel 4. 11 Hasil Klasifikasi Data Berdasarkan Threshold Mean	81
Tabel 4. 12 Hasil Pengujian <i>Threshold</i>	83
Tabel 4. 13 Hasil Perhitungan Metrik Evaluasi	91
Tabel 4. 14 Analisis Perbandingan Kinerja Berdasarkan <i>Threshold</i>	94
Tabel 4. 15 Contoh Analisis Kesalahan Klasifikasi dengan Threshold Mean.....	95

ABSTRAK

Alfatih, Syafri Maulana. 2026. Pencocokan Judul dan Transkrip Video Menggunakan Metode Cosine Similarity. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. Fachrul Kurniawan, ST., M.MT., IPU, (II) Fatchurrochman, M.Kom.

Kata kunci: Sentence BERT, Cosine Similarity, Judul Video, Transkrip Video, Kemiripan Semantik.

Perkembangan platform berbagi video menyebabkan meningkatnya jumlah konten digital yang dapat diakses oleh masyarakat, namun tidak semua judul video mampu merepresentasikan isi video secara akurat sehingga berpotensi menimbulkan kesalahpahaman informasi. Penelitian ini bertujuan untuk mengukur tingkat kecocokan antara judul dan transkrip video menggunakan metode Cosine Similarity dengan representasi semantik berbasis Sentence BERT. Dataset yang digunakan adalah *Indonesian Stand-Up Comedy Transcription Dataset* dari Mendeley Data dengan jumlah 2.774 data setelah melalui proses pembersihan data. Tahapan penelitian meliputi *text cleaning*, *case folding*, pembentukan *embedding* menggunakan model Sentence BERT *all-MiniLM-L6-v2*, serta perhitungan tingkat kemiripan menggunakan Cosine Similarity. Hasil perhitungan kemudian diklasifikasikan ke dalam kategori relevan dan tidak relevan menggunakan beberapa skenario *threshold*, yaitu 0,10; 0,15; 0,20; 0,25; dan nilai rata-rata (*mean similarity*) sebesar 0,34. Evaluasi dilakukan menggunakan Confusion Matrix dengan metrik Accuracy, Precision, Recall, dan F1-Score. Hasil penelitian menunjukkan bahwa kombinasi Sentence BERT dan Cosine Similarity mampu mengukur tingkat kecocokan antara judul dan transkrip video secara efektif. Berdasarkan hasil pengujian, *threshold* 0,34 menghasilkan performa terbaik dengan nilai Accuracy 79,82%, Precision 78,12%, Recall 82,12%, dan F1-Score 80,07%. Hasil tersebut menunjukkan bahwa representasi semantik yang dihasilkan Sentence BERT mampu memahami kesamaan makna antara judul dan transkrip video dengan baik sehingga dapat digunakan untuk mengevaluasi relevansi konten digital secara lebih objektif.

ABSTRACT

Alfatih, Syafri Maulana. 2026. Title and Video Transcript Matching Using the Cosine Similarity Method. Undergraduate Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang. Supervisors: (I) Dr. Ir. Fachrul Kurniawan, S.T., M.MT., IPU, (II) Fatchurrochman, M.Kom.

Key words: Sentence BERT, Cosine Similarity, Video Title, Video Transcript, Semantic Similarity.

The rapid growth of video sharing platforms has increased the amount of digital content accessible to the public. However, not all video titles accurately represent the content of the videos, potentially leading to misinformation and misunderstanding. This study aims to measure the level of similarity between video titles and transcripts using the Cosine Similarity method with semantic representations generated by Sentence BERT. The dataset used in this study is the Indonesian Stand-Up Comedy Transcription Dataset obtained from Mendeley Data, consisting of 2,774 records after the data cleaning process. The research stages include text cleaning, case folding, embedding generation using the Sentence BERT all-MiniLM-L6-v2 model, and similarity calculation using Cosine Similarity. The resulting similarity scores were then classified into relevant and irrelevant categories using several threshold scenarios, namely 0.10, 0.15, 0.20, 0.25, and the mean similarity value of 0.34. System evaluation was conducted using a Confusion Matrix with Accuracy, Precision, Recall, and F1-Score metrics. The results show that the combination of Sentence BERT and Cosine Similarity is effective in measuring the similarity between video titles and transcripts. Based on the experimental results, the threshold value of 0.34 achieved the best performance with an Accuracy of 79.82%, Precision of 78.12%, Recall of 82.12%, and F1-Score of 80.07%. These findings indicate that the semantic representations generated by Sentence BERT are capable of capturing the semantic similarity between video titles and transcripts, making them suitable for evaluating the relevance of digital content in a more objective manner.

البحث مستخلص

الفتاح، شافري مولانا. 2026. مطابقة العناوين ونصوص مقاطع الفيديو باستخدام طريقة تشابه جيب التمام. أطروحة البكالوريوس. برنامج دراسة هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية في فاتشورهمان، م.كوم (II)مالانغ. المشرفون: (1) د. إير. فخرول كورنياوان، س.ت.، م.م.ت.، آي.ب.يو،

التشابه الدلالي ، تشابه جيب التمام، عنوان الفيديو، نص الفيديو، Sentence BERT، لكلمات المفتاحية

أدى النمو السريع لمنصات مشاركة مقاطع الفيديو إلى زيادة كمية المحتوى الرقمي المتاح للجمهور. ومع ذلك، لا تعبر جميع عناوين مقاطع الفيديو بدقة عن محتوى هذه المقاطع، مما قد يؤدي إلى انتشار المعلومات الخاطئة وسوء الفهم. تهدف هذه الدراسة إلى قياس مستوى التشابه بين عناوين مقاطع الفيديو ونصوصها باستخدام طريقة تشابه جيب التمام مع التمثيلات الدلالية التي تم مجموعة البيانات المستخدمة في هذه الدراسة هي «مجموعة بيانات نصوص الكوميديا Sentence BERT. إنشاؤها بواسطة التي تم الحصول عليها من (Indonesian Stand-Up Comedy Transcription Dataset) «الارتجالية الإندونيسية ، وتتألف من 2,774 سجلًا بعد عملية تنقية البيانات. وتشمل مراحل البحث تنقية النص، وتحويل الأحرف Mendeley Data ، وحساب التشابه Sentence BERT all-MiniLM-L6-v2 الكبيرة والصغيرة، وإنشاء التضمين باستخدام نموذج باستخدام طريقة تشابه الجيب. ثم تم تصنيف درجات التشابه الناتجة إلى فئتين: ذات صلة وغير ذات صلة، باستخدام عدة سيناريوهات للعتبة، وهي 0.10 و 0.15 و 0.20 و 0.25، وقيمة التشابه المتوسطة البالغة 0.34. وأجري تقييم النظام باستخدام والاسترجاع (Precision) والدقة النوعية (Accuracy) مع مقاييس الدقة (Confusion Matrix) مصفوفة الارتباك (Cosine) ومقياس التشابه الجيبى Sentence BERT تُظهر النتائج أن الجمع بين نموذج F1 ودرجة (Recall) فعال في قياس التشابه بين عناوين مقاطع الفيديو ونصوصها. استنادًا إلى نتائج التجربة، حققت قيمة العتبة البالغة (Sentence Similarity) بلغت 0.34 أفضل أداء بدقة بلغت 79.82٪، ودقة تحديد بلغت 78.12٪، ومعدل استرجاع بلغ 82.12٪، ودرجة قدرة على التقاط التشابه الدلالي Sentence BERT 80.07٪. تشير هذه النتائج إلى أن التمثيلات الدلالية التي يولدها نموذج بين عناوين مقاطع الفيديو ونصوصها، مما يجعلها مناسبة لتقييم مدى صلة المحتوى الرقمي بطريقة أكثر موضوعية

BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi di masa sekarang memiliki kelebihan untuk memudahkan masyarakat dalam mengikuti perkembangan zaman, salah satunya adalah memudahkan masyarakat dalam mencari dan mendapatkan informasi. Untuk mendapatkan informasi, masyarakat biasa bergantung pada media massa sebagai sarana penerimaan informasi. Berkat teknologi yang begitu berkembang, kini telah banyak platform yang memudahkan masyarakat untuk mengakses berita atau informasi faktual yang ada. Salah satu jenis platform tersebut adalah youtube (Utami 2024). Platform ini tidak hanya berfungsi sebagai sarana hiburan, tetapi juga menjadi sumber utama dalam memperoleh informasi, pembelajaran, hingga berita. Namun, banyaknya video yang diunggah setiap harinya menimbulkan tantangan baru, yaitu bagaimana memastikan bahwa judul video benar-benar sesuai dengan isi atau transkrip video yang disajikan. Ketidaksesuaian antara judul dan transkrip video dapat menimbulkan kesalahpahaman informasi, bahkan berpotensi menyesatkan penonton dalam proses pencarian informasi yang akurat.

Masalah kecocokan antara judul dan transkrip video menjadi penting karena judul yang tidak sesuai menambahkan unsur ketidakjujuran, menggunakan bujukan yang tidak mencerminkan konten yang disampaikan seolah-olah (Susilowati 2021). Ada pembuat konten menggunakan judul yang bersifat sensasional untuk meningkatkan jumlah penonton, meskipun isi video tidak selalu sesuai dengan yang dijanjikan. Fenomena ini dikenal sebagai *clickbait*. Oleh karena itu, dibutuhkan

suatu sistem yang mampu mengukur tingkat kecocokan antara judul dan transkrip video secara objektif.

Salah satu pendekatan utama yang dapat digunakan untuk mengukur kecocokan antara judul dan transkrip video, adalah metode *Cosine Similarity*, yang mengukur kemiripan dua vektor berdasarkan sudut di antara keduanya. Nilai *cosine* yang lebih tinggi menunjukkan bahwa dua teks tersebut semakin mirip dalam arti semantik (Maulidya Prastita Syah et al. 2025). *Cosine Similarity* dapat menunjukkan sejauh mana transkrip video mendukung atau sesuai dengan judul yang digunakan. Sistem akan memanfaatkan model bahasa alami berbasis arsitektur *Transformer* yaitu *Bidirectional Encoder Representations from Transformers* (BERT) yang khusus untuk tugas pemahaman semantik kata yaitu *Sentence BERT* (S-BERT) (Febianto Agiharta, Renaldy Suteja, and Ayub Mewati 2024). *Sentence BERT* berguna untuk mengubah kalimat dalam dokumen menjadi vektor *embedding* yang menangkap makna semantik. Penggunaan model *Sentence BERT* terbukti mampu meningkatkan relevansi hasil dengan mengubah teks menjadi representasi *embedding* yang dapat diproses secara komputasi (Fathuddin, Mandyartha, and Nurlaili 2025).

Sebelum dilakukan perhitungan *Cosine Similarity* dan pembentukan representasi kalimat menggunakan *Sentence BERT*, diperlukan proses *preprocessing* untuk menyiapkan data teks agar hasil perhitungan menjadi lebih akurat. Tahapan *preprocessing* yang digunakan meliputi *text cleaning* dan *case folding*, (Kho and Tampinongkol 2025). Teks kemudian diproses menggunakan model *Sentence BERT* untuk menghasilkan representasi *embedding* berbasis

konteks semantik. *Embedding* yang dihasilkan selanjutnya digunakan dalam perhitungan *Cosine Similarity* untuk mengukur tingkat kemiripan antar teks secara lebih akurat (Rohimim 2025).

Hasil dari sistem yang dibangun dalam penelitian ini akan menghasilkan nilai tingkat kecocokan antara judul dan transkrip video yang diklasifikasikan ke dalam dua kategori, yaitu relevan dan tidak relevan. Klasifikasi ini diharapkan dapat memberikan gambaran yang jelas mengenai tingkat keakuratan representasi judul terhadap transkrip video. Sistem ini tidak hanya berguna bagi pengguna umum untuk menilai kejujuran konten, tetapi juga dapat digunakan oleh pengembang platform dalam mendeteksi potensi kesalahan informasi atau konten menyesatkan berdasarkan perhitungan teks otomatis.

Pembuatan judul yang tidak sesuai dengan isi konten termasuk perbuatan yang tidak sesuai dengan prinsip kejujuran dalam menyampaikan informasi. Allah SWT berfirman dalam Surah Al-Hujurat ayat 6,

يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ
 نَذِيرٌ ﴿٦﴾ (الحجرات/49: 6)

“Wahai orang-orang yang beriman, jika datang kepadamu seorang fasik membawa suatu berita, maka periksalah dengan teliti, agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu.” (Q.S. Al-Hujurat: 6)

Ayat ini menguraikan bagaimana berlaku dengan sesama manusia, termasuk kepada orang fasik. Diawali dengan tuntunan bagaimana menghadapi orang fasik, Allah berfirman, Wahai orang-orang yang beriman! Jika seseorang yang fasik datang kepadamu membawa suatu berita yang penting, maka ja-nganlah kamu

tergesa-gesa menerima berita itu, tetapi telitilah terlebih dahulu kebenarannya. Hal ini penting dilakukan agar kamu tidak mencelakakan suatu kaum karena kebodohan atau kecerobohan kamu mengikuti berita itu yang akhirnya kamu menyesali perbuatanmu itu yang terlanjur kamu lakukan. Ayat ini memberikan tuntunan kepada kaum muslim agar berhati-hati dalam menerima berita terutama jika bersumber dari orang yang fasik. Perlunya berhati-hati dalam menerima berita adalah untuk menghindari penyesalan akibat tindakan yang diakibatkan oleh berita yang belum diteliti kebenarannya (Kemenag 2023).

Tafsir tersebut mengingatkan pentingnya verifikasi dan kehati-hatian dalam menerima serta menyebarkan informasi. Dalam konteks penelitian ini, sistem yang mampu menilai kecocokan antara judul dan transkrip video dapat menjadi wujud penerapan nilai Islam dalam teknologi, yaitu memastikan kebenaran dan kejujuran dalam penyebaran informasi digital.

1.2 Pernyataan Masalah

Bagaimana mengukur tingkat kecocokan antara judul dan transkrip video dengan memanfaatkan representasi semantik menggunakan *Sentence BERT* dan perhitungan *Cosine Similarity*?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Penelitian ini dibatasi hanya pada proses perhitungan tingkat kecocokan antara judul dan transkrip video yang terdapat dalam dataset berbahasa Indonesia dari Mendeley Data, di mana data yang

digunakan hanya mencakup kolom Judul dan Transkrip tanpa melibatkan elemen lain seperti *author*, deskripsi, maupun *thumbnail*, sehingga fokus penelitian terpusat pada hubungan semantik antara dua komponen teks tersebut.

2. Proses pengolahan data teks pada penelitian ini dibatasi pada tahapan *preprocessing* yang meliputi *text cleaning* dan *case folding*. Data teks direpresentasikan dalam bentuk embedding menggunakan model *Sentence BERT* untuk menghasilkan representasi semantik yang lebih bermakna, dan digunakan dalam perhitungan tingkat kecocokan menggunakan metode *Cosine Similarity* sebagai pendekatan utama.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mengembangkan sistem yang mampu menghitung tingkat kecocokan antara judul dan transkrip video menggunakan metode *Cosine Similarity* dengan representasi semantik berbasis *Sentence BERT*, serta didukung oleh proses *preprocessing* yang meliputi *text cleaning* dan *case folding*. Sistem yang dihasilkan diharapkan dapat memberikan klasifikasi tingkat relevansi secara objektif ke dalam kategori relevan dan tidak relevan sebagai bentuk evaluasi kecocokan antara judul dan transkrip video.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat dalam bidang *text mining* dan *natural language processing*. Penelitian ini menghasilkan sistem yang

mampu menghitung tingkat kecocokan antara judul dan transkrip video secara objektif menggunakan representasi semantik berbasis *Sentence BERT* dan perhitungan *Cosine Similarity*. Sistem tersebut dapat dimanfaatkan sebagai pendekatan untuk mengidentifikasi tingkat kecocokan antara judul dan transkrip video, sehingga berpotensi membantu dalam mendeteksi indikasi penggunaan judul yang kurang sesuai atau bersifat *clickbait* pada platform digital. Penelitian ini juga diharapkan dapat menjadi referensi empiris bagi penelitian selanjutnya yang berkaitan dengan perhitungan kemiripan teks serta penerapan model berbasis *deep learning*, khususnya *Sentence BERT*, dalam pengukuran kecocokan antar dokumen. Penelitian ini dapat menjadi dasar pengembangan sistem perhitungan teks yang lebih efektif dalam mengevaluasi kecocokan informasi pada konten digital.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terdahulu

Untuk mengetahui efektivitas metode representasi teks dalam mengukur tingkat kecocokan antara judul dan transkrip video, sejumlah penelitian terdahulu telah memberikan kontribusi signifikan dalam bidang pengolahan teks. Beberapa penelitian yang relevan membahas penggunaan model berbasis *deep learning* seperti *Sentence BERT* dalam menghasilkan representasi semantik, pemanfaatan *Cosine Similarity* dalam pengukuran kemiripan dokumen, serta implementasi teknik *preprocessing* untuk meningkatkan akurasi pencocokan teks. Hasil dari penelitian-penelitian tersebut menunjukkan bahwa representasi berbasis *embedding* semantik yang dikombinasikan dengan pengukuran kemiripan vektor mampu memberikan hasil yang lebih efektif dalam mengidentifikasi tingkat relevansi antar dokumen. Ringkasan penelitian terdahulu yang mencakup aspek peneliti, judul, metode yang digunakan, hasil penelitian, serta perbedaan dengan penelitian ini disajikan dalam Tabel 2.1. Penelitian-penelitian tersebut menjadi landasan konseptual dalam merancang sistem pengukuran kecocokan antara judul dan transkrip video berbasis representasi semantik pada penelitian ini.

Tabel 2. 1 Penelitian Terdahulu

No	Peneliti/Judul	Metode	Hasil	Perbedaan Penelitian
1.	Fathuddin et al. (2025) Penerapan <i>Sentence BERT</i> dan <i>Cosine Similarity</i> untuk Pencarian Dokumen Skripsi	<i>Sentence BERT</i> + <i>Cosine Similarity</i> + Ontologi.	Mampu meningkatkan relevansi pencarian dokumen dibanding keyword based hingga 88%.	Fokus pada perhitungan kecocokan antara judul dan transkrip video dengan 2774 data.
2.	Holis et al. (2025) Semantic FAQ Chatbot Using SBERT(<i>Sentence BERT</i>) and <i>Cosine Similarity</i> for Academic Services	<i>Sentence BERT</i> + <i>Cosine Similarity</i>	Meningkatkan akurasi jawaban chatbot berbasis FAQ dengan 300 data akurasi 90%.	Fokus pada evaluasi kecocokan teks dengan 2774 data.
3.	Hakiki et al. (2025) Implementasi Model <i>Sentence BERT</i> untuk Deteksi Plagiarisme Pada Karya Tulis Ilmiah Psikologi Berbahasa Indonesia	<i>Sentence BERT</i> + <i>Cosine Similarity</i>	Efektif mendeteksi plagiarisme terutama copy paste dengan hasil F1 Score 87%.	Fokus pada kecocokan antara judul dan transkrip video dengan 2774 data.
4.	Fadlilah (2025) Peringkasan Teks Ekstraktif Pada Artikel Berita Bahasa Indonesia Menggunakan Metode Bert Dan <i>Cosine Similarity</i>	IndoBERT + <i>Cosine Similarity</i>	Berhasil menghasilkan ringkasan relevan (evaluasi ROUGE) dengan akurasi 78%.	Fokus pada pencocokan judul dan transkrip video menggunakan model <i>Sentence BERT</i> .
5.	Martin et al. (2025) Implementasi <i>Sentence Transformer</i> Berbasis Bert Untuk Edukasi Dan Informasi Tentang Hiv/Aids	BERT/ <i>Sentence Transformer</i>	Mampu memahami konteks pertanyaan secara semantic dengan akurasi 85%.	Menghitung kecocokan antar dokumen dengan model <i>Sentence BERT</i> dan <i>Cosine Similarity</i> .
6.	Hartatik & Syafrianto (2025) Penerapan Model	<i>Sentence BERT</i> + <i>Cosine Similarity</i>	Rekomendasi buku lebih relevan berdasarkan	Fokus pada judul dan transkrip video.

	<i>Sentence BERT</i> Untuk Sistem Rekomendasi Buku Berbasis Konten Di Perpustakaan Digital	(Content Based Filtering)	deskripsi dengan hasil precision 86%.	
7.	Febian et al. (2024) Penerapan <i>Sentence BERT</i> Untuk Similaritas Kompetensi Pekerjaan dan Mata Kuliah	<i>Sentence BERT</i>	Mampu mengukur kesamaan kompetensi secara semantic dengan akurasi 87%.	Berfokus pada dokumen judul dan transkrip video.

Berdasarkan Tabel 2.1 mengenai penelitian terdahulu, terdapat berbagai penelitian yang telah mengkaji penggunaan metode berbasis *semantic Similarity* dalam pengolahan teks. Penelitian yang dilakukan oleh (Fathuddin et al. 2025) menunjukkan bahwa penerapan *Sentence BERT* yang dikombinasikan dengan *Cosine Similarity* mampu meningkatkan relevansi pencarian dokumen skripsi, dengan hasil akurasi yang cukup tinggi dibandingkan metode berbasis kata kunci. Selanjutnya, penelitian oleh (Holis, Utomo, and Hutabarat 2025) pada sistem chatbot berbasis FAQ juga membuktikan bahwa *Sentence BERT* efektif dalam memahami konteks pertanyaan dan menghasilkan jawaban yang lebih akurat, dengan tingkat akurasi mencapai sekitar 90% pada dataset FAQ. Penelitian lain oleh (Hakiki, Atqiya, and Suryan 2025) dalam deteksi plagiarisme menunjukkan bahwa pendekatan *Sentence BERT* mampu mendeteksi kemiripan teks secara semantik dengan baik, ditunjukkan oleh nilai F1-Score yang tinggi, terutama dalam kasus *copy paste*.

Pada penelitian (Rohimim 2025) dalam penelitiannya mengenai peringkasan teks berita menggunakan BERT dan *Cosine Similarity* berhasil

menghasilkan ringkasan yang relevan dengan nilai evaluasi ROUGE yang cukup baik, meskipun masih terbatas pada domain teks berita. Penelitian oleh (Ompusunggu et al. 2025) juga menunjukkan bahwa model berbasis BERT atau *Sentence Transformer* mampu memahami hubungan antar teks secara lebih mendalam, sehingga meningkatkan kualitas pengukuran kemiripan dibandingkan metode tradisional. Menurut (Hartatik and Syafrianto 2025) dalam sistem rekomendasi buku berbasis *Sentence BERT* membuktikan bahwa penggunaan embedding semantik dapat meningkatkan relevansi rekomendasi dengan nilai precision yang tinggi. Pada penelitian oleh (Febianto Agiharta et al. 2024) menunjukkan bahwa *Sentence BERT* mampu mengukur kesamaan kompetensi secara semantik dengan akurasi yang baik, sehingga efektif digunakan dalam berbagai kasus pengukuran kemiripan teks.

Penelitian-penelitian tersebut menunjukkan bahwa penggunaan model berbasis *deep learning*, khususnya *Sentence BERT*, yang dikombinasikan dengan *Cosine Similarity* mampu memberikan hasil yang lebih akurat dalam mengukur kemiripan teks dibandingkan pendekatan berbasis frekuensi kata. Tetapi sebagian besar penelitian masih berfokus pada domain seperti pencarian dokumen, chatbot, rekomendasi, dan deteksi plagiarisme, serta menggunakan jenis data yang berbeda seperti dokumen teks, FAQ, atau deskripsi buku. Penelitian ini memiliki perbedaan pada penggunaan data berupa judul dan transkrip video dengan jumlah dataset yang lebih besar, yaitu sebanyak 2774 data, sehingga diharapkan dapat memberikan kontribusi dalam mengukur tingkat kecocokan antara judul dan isi konten video secara lebih komprehensif.

2.2 Transkrip Video

Transkrip video merupakan hasil konversi otomatis atau manual dari data audio menjadi teks yang merepresentasikan isi percakapan atau narasi dalam video. Dalam konteks penelitian media digital, transkrip video menjadi fondasi penting karena memungkinkan perhitungan berbasis teks terhadap konten audio-visual. (Adelpina, Prissa, and Supriadi 2024) menjelaskan bahwa dalam konteks penerjemahan subtitle YouTube, proses transkripsi dan translasi memiliki dampak besar terhadap kualitas representasi makna yang terkandung dalam video. Penelitian tersebut menyoroti bahwa kesalahan kecil dalam hasil transkrip dapat menyebabkan pergeseran makna yang signifikan, sehingga akurasi transkrip menjadi krusial bagi proses perhitungan konten yang berbasis teks.

Selain sebagai media untuk perhitungan semantik, transkrip video juga berfungsi sebagai data utama dalam proses *text mining*. (Larasakti 2023) dalam penelitiannya mengenai perhitungan sentimen pada komentar YouTube menjelaskan bahwa teks hasil transkripsi memungkinkan algoritma untuk mengekstraksi pola opini publik terhadap konten tertentu. Dalam penelitiannya, tahapan *preprocessing* seperti *text cleaning* dan *case folding* dilakukan untuk membersihkan data sehingga dapat menghasilkan akurasi yang lebih tinggi dalam pengklasifikasian sentiment (Kho and Tampinongkol 2025).

Transkrip video memainkan peran penting dalam penelitian yang berkaitan dengan kecocokan semantik antara judul dan isi video. Dengan mengonversi audio menjadi teks, kesamaan makna antara dua entitas linguistik tersebut dapat diukur menggunakan representasi semantik berbasis model *deep learning* seperti *Sentence*

BERT serta perhitungan *Cosine Similarity*. Dalam penelitian berbasis konten video, transkrip menjadi alat yang menghubungkan elemen audiovisual dengan bahasa yang terstruktur, menjadikan komponen kunci dalam membangun sistem perhitungan kecocokan konten secara objektif.

2.3 *Preprocessing Teks*

Preprocessing teks merupakan tahapan yang sangat penting sebelum langkah awal dimulainya suatu penelitian. *Preprocessing* teks dilakukan dengan tujuan dimana data awal diproses dengan melewati beberapa tahapan hingga data tersebut benar-benar siap untuk digunakan (Albab, P., and Fawaiq 2023). Tahapan ini mencakup proses pembersihan teks seperti, *text cleaning* yaitu menghapus simbol, tanda baca, atau angka yang tidak diperlukan, dan *case folding* yaitu mengubah semua karakter dalam teks di ubah menjadi huruf kecil atau *lowercase* (Kadzim 2025).

2.4 *Sentence BERT*

Sentence BERT (SBERT) merupakan versi khusus dari model *Bidirectional Encoder Representations from Transformers* (BERT) yang dibuat untuk memahami dan membuat representasi kalimat dalam bentuk angka yang bermakna. *Sentence BERT* menggunakan dua jaringan identik yang bekerja sama dengan teknik pembelajaran triplet untuk memastikan bahwa kalimat-kalimat yang memiliki makna serupa diwakili oleh angka-angka yang juga serupa. *Sentence BERT* bisa digunakan dalam mengukur seberapa mirip makna antar kalimat (Ferdiansyah 2025). Berbeda dengan BERT standar yang memerlukan proses komputasi

berulang untuk membandingkan dua kalimat, *Sentence BERT* menggunakan arsitektur *siamese network* atau *triplet network* sehingga mampu menghasilkan embedding kalimat yang dapat dibandingkan secara langsung menggunakan ukuran kemiripan seperti *Cosine Similarity*.

Sentence BERT mengubah teks menjadi representasi numerik berdimensi tinggi yang merepresentasikan makna dari keseluruhan kalimat, bukan hanya berdasarkan kemunculan kata. Hal ini memungkinkan pengukuran kemiripan teks dilakukan secara lebih akurat meskipun terdapat perbedaan kosakata atau struktur kalimat. *Sentence BERT* banyak digunakan dalam berbagai tugas *natural language processing* seperti pencarian semantik (*semantic search*), deteksi plagiarisme, sistem rekomendasi, serta pengukuran kesamaan teks (Widianoro, Dwiyooga Mustafid Sanjaya, and Ridwan 2024). Pada penelitian ini, *Sentence BERT* digunakan untuk merepresentasikan judul dan transkrip video ke dalam bentuk embedding, yang kemudian digunakan dalam perhitungan *Cosine Similarity* untuk menentukan tingkat kecocokan antara keduanya.

Keunggulan utama *Sentence BERT* terletak pada kemampuannya dalam menghasilkan embedding yang bersifat kontekstual dan dapat dibandingkan secara langsung tanpa perlu melakukan perhitungan ulang terhadap seluruh pasangan kalimat (Maulidya Prastita Syah et al. 2025). Hal ini menjadikan *Sentence BERT* lebih efisien dalam skenario dengan jumlah data yang besar, seperti pada penelitian ini yang menggunakan dataset sebanyak 2774 data. Dengan memanfaatkan representasi semantik yang dihasilkan, sistem dapat mengidentifikasi kesamaan makna secara lebih mendalam, termasuk pada kasus di mana kata yang digunakan

berbeda tetapi memiliki arti yang serupa. Penggunaan *Sentence BERT* dalam penelitian ini diharapkan mampu meningkatkan akurasi dalam mengukur tingkat kecocokan antara judul dan transkrip video dibandingkan pendekatan tradisional.

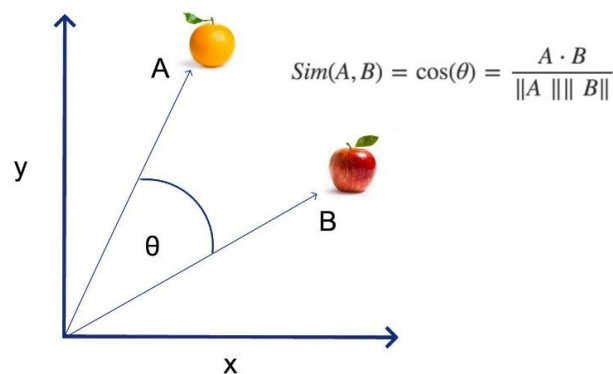
2.5 *Cosine Similarity*

Cosine Similarity merupakan salah satu metode pengukuran kesamaan teks yang paling populer dalam penelitian di Indonesia karena kesederhanaannya dan kemampuannya dalam mengukur kedekatan semantik antar dokumen. Metode ini menghitung sudut kosinus antara dua vektor teks dan menghasilkan nilai antara 0 hingga 1, dimana nilai yang lebih tinggi menunjukkan tingkat kesamaan yang lebih besar. Penelitian oleh (Lahitani et al. 2022) membuktikan efektivitas *Cosine Similarity* dalam menghitung teks di media sosial. Hasil perhitungan menunjukkan bahwa *Cosine Similarity* mampu mengidentifikasi kelompok teks dengan kemiripan tinggi secara efisien.

Dengan memanfaatkan representasi semantik berbasis *Sentence BERT* dan pengukuran kemiripan menggunakan *Cosine Similarity*, sistem mampu melakukan pengukuran kesamaan makna secara lebih kontekstual dan akurat, terutama dalam sistem yang menuntut kecepatan dan presisi dalam pencocokan teks. Perhitungan *Cosine Similarity* dilakukan untuk mengukur tingkat kesamaan antara judul dan transkrip video. Proses ini menggunakan *embedding* yang dihasilkan oleh *Sentence BERT* untuk merepresentasikan setiap teks dalam bentuk vektor numerik berdimensi tinggi. Judul dan transkrip video diubah menjadi representasi vektor yang menangkap makna keseluruhan kalimat, sehingga memungkinkan

pengukuran kemiripan dilakukan tidak hanya berdasarkan kesamaan kata, tetapi juga berdasarkan kesamaan konteks dan makna yang terkandung di dalamnya.

Cosine Similarity



Gambar 2. 1 Konsep Dasar *Cosine Similarity*

$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} \quad (2.1)$$

Keterangan:

$Sim(A, B)$: Nilai kemiripan antara vektor A dan vektor B

A : Vektor pertama (dalam penelitian ini merepresentasikan teks judul)

B : Vektor kedua (dalam penelitian ini merepresentasikan teks transkrip)

$A \cdot B$: Dot product (perkalian titik) antara vektor A dan B

$||A||$: Panjang (magnitudo) vektor A

$||B||$: Panjang (magnitudo) vektor B

θ (theta) : Sudut yang terbentuk antara kedua vector

Gambar 2.1 di atas menunjukkan ilustrasi konsep dasar *Cosine Similarity*, di mana dua vektor A dan B membentuk sudut θ yang menjadi dasar pengukuran kesamaan antar teks. Semakin kecil sudut θ , semakin besar nilai $\cos(\theta)$ yang

dihasilkan, menandakan kedua teks semakin mirip secara semantik. Pada penelitian ini, proses perhitungan *Cosine Similarity* dimulai dari representasi teks dalam bentuk vektor menggunakan embedding yang dihasilkan oleh *Sentence BERT*. Judul dan transkrip video diubah menjadi representasi numerik berdimensi tinggi yang mencerminkan makna keseluruhan kalimat.

Selanjutnya dilakukan perhitungan perkalian titik (*dot product*) antara dua vektor untuk mengukur tingkat kesamaan di antara keduanya. Lalu dihitung panjang masing-masing vektor (*magnitude*) yang diperoleh dari akar kuadrat jumlah kuadrat setiap elemen vektor embedding. Nilai *Cosine Similarity* kemudian diperoleh dengan membagi hasil perkalian titik dengan hasil kali panjang kedua vektor. Nilai akhir berkisar antara 0 hingga 1, di mana nilai mendekati 1 menunjukkan tingkat kesamaan yang tinggi atau relevan, sedangkan nilai mendekati 0 menunjukkan bahwa kedua teks memiliki perbedaan makna yang signifikan.

2.6 Metrik Evaluasi Klasifikasi

Metrik evaluasi digunakan untuk mengukur kinerja model klasifikasi dalam mengidentifikasi relevansi antara judul dan transkrip video. Pada penelitian ini, evaluasi dilakukan menggunakan Confusion Matrix yang menghasilkan empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Berdasarkan komponen tersebut dihitung nilai Accuracy, Precision, Recall, dan F1-Score untuk mengetahui performa model secara menyeluruh.

2.6.1 *Confusion Matrix*

Confusion Matrix merupakan metode evaluasi yang digunakan untuk melihat performa sistem klasifikasi dengan membandingkan hasil prediksi sistem dengan label sebenarnya (Dawud 2025). *Confusion Matrix* menyajikan informasi mengenai jumlah prediksi yang benar maupun salah yang dihasilkan oleh sistem dalam bentuk tabel matriks. *Confusion Matrix* terdiri dari empat komponen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Habibi and Widana 2025).

Tabel 2. 2 Confusion Matrix

<i>Actual/Predicted</i>	Relevan	Tidak relevan
Relevan	<i>True Positive</i> (TP)	<i>False Negative</i> (FN)
Tidak relevan	<i>False Positive</i> (FP)	<i>True Negative</i> (TN)

Keterangan:

True Positive (TP) : Jumlah data yang sebenarnya relevan dan diprediksi relevan oleh sistem

True Negative (TN) : Jumlah data yang sebenarnya tidak relevan dan diprediksi tidak relevan oleh sistem

False Positive (FP) : Jumlah data yang sebenarnya tidak relevan tetapi diprediksi relevan oleh sistem

False Negative (FN) : Jumlah data yang sebenarnya relevan tetapi diprediksi tidak relevan oleh system

Evaluasi kinerja sistem dilakukan menggunakan *Confusion Matrix* yang terdiri dari nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Nilai tersebut dihitung metrik evaluasi berupa *Accuracy*,

Precision, *Recall*, *F1-Score* untuk mengukur performa sistem dalam mengklasifikasikan relevansi antara judul dan transkrip video.

2.6.1 *Accuracy*

Accuracy merupakan metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan sistem dalam melakukan klasifikasi secara keseluruhan (Dawud 2025). Nilai *accuracy* menunjukkan proporsi jumlah prediksi yang benar dibandingkan dengan keseluruhan data yang diuji. Persamaan yang digunakan untuk menghitung evaluasi *Accuracy* ditunjukkan pada Persamaan 2.2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

Keterangan:

TP : *True Positive*

TN : *True Negative*

FP : *False Positive*

FN : *False Negative*

2.6.1 *Precision*

Precision merupakan salah satu metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan sistem dalam mengklasifikasikan data ke dalam kelas positif. Metrik ini menunjukkan seberapa banyak data yang diprediksi sebagai kelas positif oleh sistem benar-benar merupakan data yang positif berdasarkan label sebenarnya. *Precision* digunakan untuk mengevaluasi seberapa akurat sistem dalam mengidentifikasi data yang relevan dibandingkan dengan seluruh data yang diprediksi relevan berdasarkan perhitungan *Sentence BERT* dan *Cosine Similarity*.

Nilai *precision* diperoleh dengan membandingkan jumlah data yang diprediksi benar sebagai relevan (*True Positive*) dengan jumlah seluruh data yang diprediksi sebagai relevan, baik yang benar maupun yang salah. Persamaan yang digunakan untuk menghitung evaluasi *Precision* ditunjukkan pada Persamaan 2.3.

$$Precision = \frac{TP}{TP+FP} \quad (2.3)$$

Keterangan:

TP : *True Positive*

FP : *False Positive*

2.6.1 Recall

Recall menggambarkan kemampuan model dalam menemukan semua kasus positif yang sebenarnya (Nisfi et al. 2025). *Recall* digunakan untuk mengukur kemampuan sistem dalam mengidentifikasi seluruh data yang termasuk dalam kategori relevan berdasarkan hasil perhitungan *Sentence BERT* dan *Cosine Similarity*. Nilai *recall* diperoleh dengan membandingkan jumlah data yang diprediksi benar sebagai relevan (*True Positive*) dengan jumlah seluruh data yang sebenarnya relevan, baik yang berhasil diprediksi dengan benar maupun yang tidak berhasil terdeteksi oleh sistem. Persamaan yang digunakan untuk menghitung evaluasi *Recall* ditunjukkan pada Persamaan 3.4.

$$Recall = \frac{TP}{TP+FN} \quad (2.4)$$

Keterangan:

TP : *True Positive*

FN : *False Negative*

2.6.1 *F1-Score*

F1-Score merupakan metrik evaluasi yang digunakan untuk mengukur keseimbangan antara nilai *precision* dan *recall* dalam sistem klasifikasi. Pada penelitian ini, *F1-Score* digunakan untuk mengevaluasi performa sistem dalam mengklasifikasikan tingkat kecocokan antara judul dan transkrip berdasarkan perhitungan *Sentence BERT* dan *Cosine Similarity*. Nilai *F1-Score* diperoleh dari kombinasi *precision* dan *recall*, dengan rentang nilai 0 hingga 1. Semakin mendekati 1, semakin baik performa klasifikasi sistem. Persamaan perhitungan *F1-Score* ditunjukkan pada Persamaan 2.5.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.5)$$

Keterangan:

Precision : nilai ketepatan prediksi data relevan.

Recall : kemampuan sistem menemukan seluruh data relevan

BAB III

METODOLOGI PENELITIAN

Bab ini menjelaskan secara sistematis metodologi yang digunakan dalam penelitian ini untuk mencocokkan atau mencari relevansi antara judul dengan transkrip video. Metodologi penelitian ini mencakup keseluruhan tahapan yang dilakukan mulai dari pengumpulan data hingga implementasi metode *Cosine Similarity*. Tahapan-tahapan tersebut meliputi pengumpulan data dokumen yang akan menjadi dataset, perancangan sistem yang akan dibangun, preprocessing data untuk membersihkan dan mengolah teks mentah menjadi format yang dapat diproses, serta representasi teks menggunakan model *Sentence BERT* untuk menghasilkan embedding berbasis semantik. Embedding tersebut kemudian digunakan dalam penerapan metode *Cosine Similarity* untuk mengukur tingkat kemiripan antara judul dengan transkrip video. Setiap tahapan dirancang secara berurutan dan saling berkaitan untuk menghasilkan klasifikasi yang akurat.

3.1 Sumber Data

Pada penelitian ini, pengumpulan data dan informasi menggunakan data sekunder berupa *Indonesian Stand-Up Comedy Transcription Dataset* (Supriyono Supriyono; Aji Prasetya Wibawa; Suyono Suyono; Fachrul Kurniawan 2025). Berdasarkan deskripsi pada repositori tersebut, dataset ini berisi transkrip video stand-up comedy berbahasa Indonesia yang didapat dari video *Indonesian stand-up comedy* yang dipublikasikan melalui kanal YouTube Kompas TV.

Jumlah data awal yang diperoleh dari dataset tersebut sebanyak 3713 entri. Sebelum digunakan dalam proses perhitungan, dilakukan tahap seleksi dan pembersihan data (*data cleaning*) untuk memastikan kualitas data tetap terjaga. Proses pembersihan meliputi penghapusan data yang memiliki nilai kosong (*null*), dan penghilangan data duplikat. Setelah melalui tahap pembersihan, jumlah data yang layak digunakan dalam penelitian ini menjadi sebanyak 2774 entri.

Data yang telah dibersihkan kemudian difokuskan pada dua atribut utama, yaitu judul dan transkrip video, yang selanjutnya diproses melalui tahapan preprocessing sebelum dilakukan pembobotan menggunakan metode *Sentence BERT* dan pengukuran tingkat kecocokan menggunakan *Cosine Similarity*.

Penelitian ini tidak melakukan pengambilan data secara langsung dari platform YouTube, melainkan menggunakan dataset sekunder yang telah tersedia secara publik melalui Mendeley Data sebagai sumber penelitian. Berikut contoh data yang telah diperoleh:

Tabel 3. 1 Contoh Data

Judul	Transkrip
Yudha Keling Pasrah Jadi Bahan Ejekan Babe Cabita dan Uus - CALLBACK SUCI 4	hati-hati orang itu semenjak di ada di kampung kita panen kita berkurang lh itu kan orang yang digosipin itu Iya Yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin Hidayah
Selamat untuk 10 Komika yang Lolos ke Babak Grand Final! SUCI Liga Komunitas	Hai hihihi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak Indonesia

Transkrip pada tabel ditampilkan dalam bentuk cuplikan untuk menjaga keterbacaan. Teks transkrip lengkap tetap digunakan dalam proses perhitungan dan perhitungan *Similarity*.

3.2 Labelisasi Data

Proses pelabelan data dalam penelitian ini dilakukan berdasarkan tingkat kecocokan antara judul video dan transkrip video menggunakan pendekatan *pseudo labelling* atau pelabelan secara otomatis (Ayuningtyas and Yustanti 2024). Penentuan label dilakukan secara otomatis dengan memanfaatkan metode TF IDF dan perhitungan *Cosine Similarity* sebagai dasar dalam mengukur tingkat kemiripan awal antara judul dan transkrip. Nilai kemiripan yang dihasilkan kemudian digunakan sebagai acuan untuk menentukan kategori relevan dan tidak relevan berdasarkan ambang batas (*threshold*) tertentu. Data dikategorikan sebagai relevan apabila nilai kemiripan melebihi *threshold* yang ditetapkan, sedangkan data dengan nilai kemiripan di bawah *threshold* dikategorikan sebagai tidak relevan.

Proses pseudo labeling terhadap 2774 data menghasilkan 1369 data berlabel relevan dan 1405 data berlabel tidak relevan. Distribusi ini relatif seimbang sehingga layak digunakan sebagai dasar pengujian dan evaluasi sistem. Contoh hasil pelabelan ditampilkan pada Tabel 3.2, yang memuat judul video, sebagian transkrip, label, dan jumlah data per kategori.

Tabel 3. 2 Contoh Data Label

No	Judul	Transkrip	Label	Jumlah
1	Yudha Keling Pasrah Jadi Bahan Ejekan Babe Cabita dan Uus - CALLBACK SUCI 4	hati-hati orang itu semenjak di ada di kampung kita panen kita berkurang lh itu kan orang yang digosipin itu Iya Yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin Hidayah	0	1405
2	Selamat untuk 10 Komika yang Lolos ke Babak Grand Final!	Hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung	1	1369

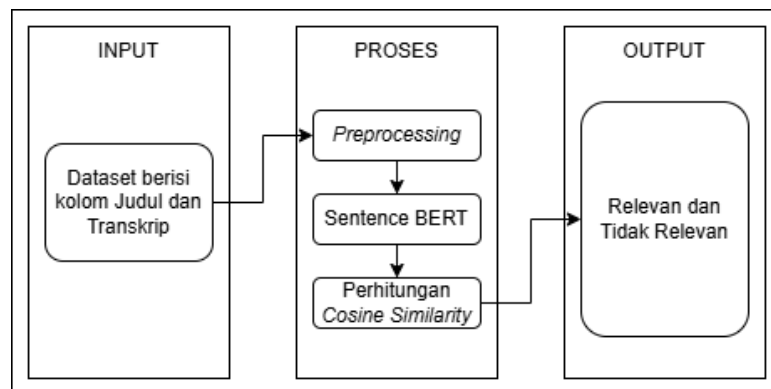
	SUCI Liga Komunitas	kita persilahkan karena kita akan langsung masuk kebabak Indonesia		
Total				2774

Tabel 3.2 menampilkan contoh data sampel yang telah melalui proses pelabelan menggunakan pendekatan *pseudo labeling*. Setiap data terdiri dari kolom judul video, sebagian transkrip, label, serta jumlah data pada masing-masing kategori. Pemberian label dilakukan secara otomatis berdasarkan nilai kemiripan antara judul dan transkrip yang dihitung menggunakan metode TF IDF dan *Cosine Similarity*. Nilai kemiripan tersebut kemudian dibandingkan dengan ambang batas (*threshold*) yang telah ditentukan untuk mengelompokkan data ke dalam kategori relevan dan tidak relevan. Label bernilai 1 menunjukkan bahwa judul dan transkrip memiliki tingkat kemiripan yang tinggi sehingga dikategorikan sebagai relevan, sedangkan label bernilai 0 menunjukkan tingkat kemiripan yang rendah sehingga dikategorikan sebagai tidak relevan.

Perlu diperhatikan bahwa transkrip yang ditampilkan pada tabel hanya merupakan sebagian dari keseluruhan transkrip. Hal ini dilakukan karena panjang transkrip pada dataset relatif besar sehingga tidak memungkinkan untuk ditampilkan secara penuh dalam tabel. Contoh data pada tabel ini bertujuan untuk memberikan gambaran mengenai proses pelabelan yang dilakukan sebelum tahap pengukuran kemiripan menggunakan metode utama, yaitu *Sentence BERT* dan *Cosine Similarity*.

3.3 Perancangan Sistem

Perancangan sistem merupakan tahap yang penting dalam suatu penelitian karena tahap ini menentukan bagaimana sistem akan dirancang dan diimplementasikan. Dalam perancangan sistem, diperlukan kejelasan dalam memahami kebutuhan sistem dan memastikan sistem yang dikelola dapat bekerja serta menghasilkan hasil yang akurat. Pada perancangan sistem melewati beberapa tahapan mekanisme yaitu input, proses, dan output dari sistem yang akan dibuat sehingga memberikan hasil pencocokan judul dengan transkrip video. Perancangan sistem yang akan dibuat dalam penelitian dapat dilihat pada Gambar 3.1.



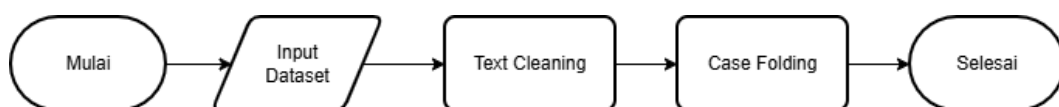
Gambar 3. 1 Perancangan Sistem

Berdasarkan Gambar 3.1 di atas, pada bagian *input*, sistem menerima dataset yang terdiri dari beberapa atribut, yaitu label, judul, dan transkrip video. Pada bagian *proses*, tahapan dimulai dengan preprocessing untuk membersihkan dan menyiapkan teks agar dapat diproses dengan baik. Kemudian teks direpresentasikan dalam bentuk embedding menggunakan model *Sentence BERT* untuk menghasilkan representasi semantik dari judul dan transkrip video. *Embedding* yang dihasilkan kemudian digunakan dalam perhitungan *Cosine Similarity* untuk mengukur tingkat

kesamaan antara keduanya. Hasil akhir dari proses tersebut ditampilkan pada bagian *output*, berupa hasil pencocokan antara judul dan transkrip video dalam bentuk klasifikasi, yaitu relevan dan tidak relevan. Klasifikasi ini diperoleh berdasarkan nilai kemiripan yang dihasilkan dari perhitungan *Cosine Similarity*, di mana nilai tersebut berada pada rentang 0 hingga 1. Nilai yang mendekati 1 menunjukkan tingkat kesamaan yang tinggi atau relevan, sedangkan nilai yang mendekati 0 menunjukkan tingkat kesamaan yang rendah atau tidak relevan. Dengan demikian, sistem mampu memberikan interpretasi yang jelas dan terukur terhadap tingkat kecocokan antara judul dan transkrip video.

3.4 *Preprocessing*

Preprocessing teks merupakan langkah awal penting untuk memastikan data bersih, konsisten, dan siap hitung. Tahapan meliputi *text cleaning* yaitu menghapus simbol, tanda baca, dan angka yang tidak diperlukan, dan *case folding* yaitu mengubah semua karakter dalam teks di ubah menjadi huruf kecil atau *lowercase* (Kadzim 2025). Hasil *preprocessing* digunakan pada model *Sentence BERT* dan pengukuran kesamaan *Cosine Similarity*. Tahapan lengkap *preprocessing* dapat dilihat pada Gambar 3.2 di bawah ini.



Gambar 3. 2 Diagram *Preprocessing*

3.4.1 Text Cleaning

Cleaning adalah proses membersihkan data dengan menghilangkan karakter-karakter yang tidak diperlukan seperti tanda baca, angka, *link*, *hashtag*, *mention*, emoji, dan lain-lain. Tujuannya adalah untuk menghilangkan *noise* dalam data teks (Nurwanda, Suarna, and Prihartono 2024). Hasil proses *case folding* dapat dilihat pada tabel 3.3 dan 3.4 berikut

Tabel 3. 3 Hasil *Cleaning* Q1 dan D1

	Judul (Q1)	Transkrip (D1)
Sebelum	Yudha Keling Pasrah Jadi Bahan Ejekan Babe Cabita dan Uus - CALLBACK SUCI 4	hatihati orang itu semenjak di ada di kampung kita panen kita berkurang lh itu kan orang yang digosipin itu Iya Yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin Hidayah
Sesudah	Yudha Keling Pasrah Jadi Bahan Ejekan Babe Cabita dan Uus CALLBACK SUCI	hatihati orang itu semenjak di ada di kampung kita panen kita berkurang ih itu kan orang yang digosipin itu iya yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin hidayah

Tabel 3. 4 Hasil *Cleaning* Q2 dan D2

	Judul (Q2)	Transkrip (D2)
Sebelum	Selamat untuk 10 Komika yang Lolos ke Babak Grand Final! SUCI Liga Komunitas	Hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak Indonesia
Sesudah	Selamat untuk Komika yang Lolos ke Babak Grand Final SUCI Liga Komunitas	Hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak Indonesia

3.4.2 Case Folding

Tahapan *case folding* merupakan proses awal dalam *preprocessing* teks yang bertujuan untuk menyeragamkan bentuk penulisan huruf pada data teks. Pada tahap ini, seluruh karakter dalam teks diubah menjadi huruf kecil (*lowercase*) agar sistem tidak membedakan kata yang sama namun memiliki perbedaan kapitalisasi.

Langkah ini penting untuk memastikan konsistensi data sehingga setiap kata memiliki representasi yang seragam dalam proses perhitungan berikutnya. Selain itu, pada tahap ini juga dilakukan penghapusan tanda baca, angka, dan karakter non alfabet yang tidak relevan dengan proses perhitungan teks. Hasil case folding akan menghasilkan teks yang lebih bersih dan siap untuk diproses ke tahap selanjutnya. Hasil proses *case folding* dapat dilihat pada tabel 3.5 dan 3.6 berikut.

Tabel 3. 5 Hasil *Case Folding* Q1 dan D1

	Judul (Q1)	Transkrip (D1)
Sebelum	Yudha Keling Pasrah Jadi Bahan Ejekan Babe Cabita dan Uus CALLBACK SUCI	hati-hati orang itu semenjak di ada di kampung kita panen kita berkurang lh itu kan orang yang digosipin itu Iya Yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin Hidayah
Sesudah	yudha keling pasrah jadi bahan ejekan babe cabita dan uus callback suci	hati-hati orang itu semenjak di ada di kampung kita panen kita berkurang ih itu kan orang yang digosipin itu iya yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin hidayah

Tabel 3. 6 Hasil *Case Folding* Q2 dan D2

	Judul (Q2)	Transkrip (D2)
Sebelum	Selamat untuk Komika yang Lolos ke Babak Grand Final SUCI Liga Komunitas	Hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak Indonesia
Sesudah	selamat untuk komika yang lolos ke babak grand final suci liga komunitas	hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak indonesia

3.5 Perhitungan *Sentence BERT*

Sentence BERT digunakan dalam penelitian ini untuk merepresentasikan teks judul dan transkrip video ke dalam bentuk embedding berbasis semantik. Berbeda dengan pendekatan berbasis frekuensi kata, penggunaan *Sentence BERT* memungkinkan setiap teks diubah menjadi vektor numerik yang merepresentasikan

makna keseluruhan kalimat. Proses ini bertujuan agar pengukuran kemiripan antara judul dan transkrip dapat dilakukan secara lebih kontekstual. Setiap teks yang telah melalui *preprocessing* diproses menggunakan model *Sentence BERT* untuk menghasilkan *embedding* kalimat. *Embedding* tersebut kemudian digunakan sebagai input dalam perhitungan *Cosine Similarity* untuk menentukan tingkat kecocokan antara judul dan transkrip video. Pada penelitian ini tidak dilakukan proses *fine tuning* terhadap model *Sentence BERT*. Hal ini dikarenakan model yang digunakan merupakan *pretrained* model yaitu *all-MiniLM-L6-v2* yang telah dilatih pada dataset skala besar dan dirancang untuk menghasilkan representasi semantik kalimat yang efektif dalam tugas pengukuran kecocokan teks (Alwasi'i 2025). Penelitian ini berfokus pada pengukuran tingkat kemiripan teks menggunakan metode *Cosine Similarity*, bukan pada proses pelatihan atau optimasi model. Sistem tidak hanya mempertimbangkan kesamaan kata, tetapi juga kesamaan makna, sehingga diharapkan mampu meningkatkan akurasi dalam proses klasifikasi relevansi.

3.5.1 Tokenisasi *Sentence BERT*

Proses ini dilakukan menggunakan *tokenizer* yang sesuai dengan model *Sentence BERT* yang digunakan yaitu *WordPiece Tokenizer*. Tokenisasi bertujuan untuk memecah kalimat menjadi unit token yang lebih kecil sehingga dapat diproses oleh model. Setelah teks diubah menjadi token, model *Sentence BERT* digunakan untuk menghasilkan representasi vektor dari token-token tersebut (Maulidya Prastita Syah et al. 2025). Berikut adalah contoh tokenisasi *Sentence BERT* pada Q1 D1 dan Q2 D2.

Tabel 3. 7 Tokenisasi *Sentence BERT* Q1 dan D1

	Judul (Q2)	Transkrip (D2)
Sebelum	yudha keling pasrah jadi bahan ejekan babe cabita dan uus callback suci	hati-hati orang itu semenjak di ada di kampung kita panen kita berkurang ih itu kan orang yang digosipin itu iya yang kalau misalnya pakai baju bajunya dilepas dirubung lalat tetapi itu mungkin hidayah
Sesudah	['yu', '##dha', 'ke', '##ling', 'pas', '##rah', 'ja', '##di', 'baha', '##n', 'e', '##je', '##k', 'babe', 'cab', '##ita', 'u', '##us', 'call', '##back', 'su', '##ci']	['hat', '##ih', '##ati', 'or', '##ang', 'se', '##men', '##jak', 'kam', '##pu', '##ng', 'pan', '##en', 'ku', '##rang', 'i', '##h', 'kan', 'or', '##ang', 'dig', '##osi', '##pin', 'i', '##ya', 'kala', '##u', 'mis', '##al', 'pak', '##ai', 'ba', '##ju', 'ba', '##ju', 'le', '##pas', 'rub', '##ung', 'lal', '##at', 'mu', '##ng', '##kin', 'hid', '##aya', '##h', '.....]

Tabel 3. 8 Tokenisasi *Sentence BERT* Q2 dan D2

	Judul (Q2)	Transkrip (D2)
Sebelum	selamat untuk komika yang lolos ke babak grand final suci liga komunitas	hai hihhi terhenti dan kita akan memanggil pesertapeserta sebelumnya untuk naik keatas panggung kita persilahkan karena kita akan langsung masuk kebabak indonesia
Sesudah	['se', '##lam', '##at', 'ko', '##mi', '##ka', 'lo', '##los', 'baba', '##k', 'grand', 'final', 'su', '##ci', 'liga', 'ko', '##mun', '##itas']	['hai', 'hi', '##hi', '##hi', 'hen', '##ti', 'pang', '##gil', 'pe', '##ser', '##ta', '##pes', '##ert', '##a', 'bel', '##um', 'na', '##ik', 'ata', '##s', 'pang', '##gun', '##g', 'si', '##lah', 'lang', '##sun', '##g', 'mas', '##uk', 'baba', '##k', 'indonesia', '.....]

Tokenisasi pada *Sentence BERT* merupakan proses internal dari model yang digunakan untuk mengubah teks menjadi subword token. Proses ini tidak dilakukan secara manual, melainkan otomatis oleh tokenizer bawaan model. Token yang memiliki awalan “##” menunjukkan bahwa token tersebut merupakan bagian dari *subword tokenization* yang digunakan oleh tokenizer *Sentence BERT*, yaitu metode pemecahan kata menjadi bagian-bagian yang lebih kecil agar dapat diproses oleh model secara lebih efektif. Sebuah kata tidak selalu dipertahankan sebagai satu token utuh, melainkan dapat dipecah menjadi beberapa bagian kata (Devlin et al. 2019). Tanda “##” menandakan bahwa token tersebut merupakan kelanjutan dari token sebelumnya dan bukan awal dari kata baru.

3.5.2 Token *Embedding*

Token Embedding yaitu proses mengubah kalimat-kalimat yang telah dibersihkan menjadi representasi vektor menggunakan model *Sentence BERT*. Vektor *embedding* ini merepresentasikan makna dari setiap kata dalam bentuk angka untuk menghasilkan vektor *embedding* berdimensi 384 (Fathuddin et al. 2025). Berikut adalah contoh hasil embedding menggunakan *Sentence BERT* pada data judul dan transkrip video pada Q1 D1 dan Q2 D2.

Tabel 3. 9 Token *Embedding Sentence BERT* pada Q1

Token	dim1	dim2	dim3	dim4	 dim384
[CLS]	-0,38582	-0,16416	-0,21854	-0,07298	
yu	-0,50627	-0,21964	-0,07564	-0,11855	
##dha	-0,18193	-0,11452	-0,17128	-0,03196	
ke	-0,38178	0,181362	0,286886	0,232073	
##ling	-0,80506	0,008877	0,109486	0,423132	
.....					

Tabel 3. 10 Token *Embedding Sentence BERT* pada D1

Token	dim1	dim2	dim3	dim4	 dim384
[CLS]	0,134951	0,343718	0,149452	-0,23348	
hat	0,110522	0,197401	0,186959	-0,27697	
##ih	0,056213	0,267403	0,182162	-0,01081	
##ati	0,457862	0,399329	-0,10798	0,083305	
or	-0,16726	0,436802	-0,10951	-0,14204	
.....					

Tabel 3. 11 Token *Embedding Sentence BERT* pada Q2

Token	dim1	dim2	dim3	dim4	 dim384
[CLS]	0,044291	-0,01574	0,185095	-0,34668	
se	-0,76227	-0,08399	0,292581	-0,09425	
##lam	-0,21587	0,082334	0,045041	-0,05954	
##at	-0,0606	0,61311	-0,17889	-0,18959	
ko	-0,19563	0,223572	0,001307	-0,01214	
.....					

Tabel 3. 12 Token *Embedding Sentence BERT* pada D2

Token	dim1	dim2	dim3	dim4	 dim384
[CLS]	0,038238	0,031355	0,15423	-0,33902	
hai	0,013802	-0,28037	0,165946	-0,10889	
hi	-0,14259	0,22964	0,316572	-0,23666	

##hi	0,006247	0,153506	0,055443	0,024941	
##hi	0,183861	0,162828	-0,06932	0,053162	
.....					

Nilai *embedding* yang ditampilkan pada tabel hanya sebagian dari keseluruhan dimensi (384 dimensi) yang dihasilkan oleh model *Sentence BERT*. Seluruh nilai *embedding* digunakan dalam proses perhitungan *Cosine Similarity*.

3.5.3 Mean Pooling

Mean pooling merupakan tahapan dalam proses *Sentence BERT* yang digunakan untuk mengubah *embedding* pada tingkat token menjadi satu vektor representasi kalimat (*sentence embedding*). Setelah teks melalui proses tokenisasi dan *encoding*, model *transformer* menghasilkan *embedding* untuk setiap token yang mencerminkan makna kontekstual dari masing-masing token. Untuk keperluan pengukuran kemiripan antar teks, diperlukan satu vektor yang dapat merepresentasikan keseluruhan kalimat. Digunakan metode *mean pooling* untuk menggabungkan seluruh *embedding* token menjadi satu vektor tunggal.

Proses *mean pooling* dilakukan dengan menghitung rata-rata dari seluruh *embedding* token dalam satu kalimat. Dengan pendekatan ini, informasi dari setiap token dapat dirangkum sehingga menghasilkan representasi semantik yang utuh dan stabil. Metode ini umum digunakan dalam *Sentence BERT* karena mampu menghasilkan representasi kalimat yang lebih representatif dibandingkan hanya menggunakan satu token tertentu. Persamaan yang digunakan untuk menghitung *mean pooling* ditunjukkan pada Persamaan 3.1.

$$v = \frac{1}{n} \sum_{i=1}^n h_i \quad (3.1)$$

Keterangan:

v : Vektor representasi kalimat

n : Jumlah token dalam satu kalimat

h_i : Vektor embedding token ke- i

Hasil dari proses mean pooling berupa vektor representasi kalimat atau *sentence embedding* yang kemudian ditampilkan pada Tabel 3.19 dan Tabel 3.20.

Tabel 3. 13 *Mean Pooling Sentence BERT* Q1 dan D1

<i>Embedding</i>	dim1	dim2	dim3	dim4	 dim384
Judul (Q1)	-0,36716	0,07133	-0,13159	0,043961	
Transkrip (D1)	0,051342	0,148609	0,056688	-0,04716	

Tabel 3. 14 *Mean Pooling Sentence BERT* Q2 dan D2

<i>Embedding</i>	dim1	dim2	dim3	dim4	 dim384
Judul (Q2)	-0,12524	0,216835	0,05115	-0,06025	
Transkrip (D2)	0,050575	-0,05165	0,018391	-0,03466	

Hasil yang ditampilkan pada Tabel 3.13 dan Tabel 3.14 merupakan vektor representasi kalimat yang diperoleh dari proses *mean pooling* terhadap *embedding* setiap token. Proses ini menghasilkan satu vektor yang merepresentasikan keseluruhan makna kalimat secara semantik. Setiap nilai pada dimensi vektor mencerminkan fitur semantik tertentu yang dihasilkan oleh model *Sentence BERT*. Vektor yang dihasilkan memiliki dimensi tinggi, yaitu sebesar 384 dimensi, namun pada tabel hanya ditampilkan sebagian dimensi sebagai representasi. Seluruh nilai vektor tersebut digunakan dalam proses perhitungan *Cosine Similarity* untuk mengukur tingkat kemiripan antara judul dan transkrip video. Perbedaan nilai pada

setiap dimensi menunjukkan adanya perbedaan representasi semantik antara kedua teks yang dibandingkan.

3.6 *Cosine Similarity*

Cosine Similarity merupakan salah satu metode pengukuran kesamaan antar dua vector metode dari dokumen tersebut (Maulidya Prastita Syah et al. 2025). Metode ini menghitung nilai kosinus dari sudut yang terbentuk antara dua vektor dokumen dalam ruang multidimensi. Semakin kecil sudut yang terbentuk antara kedua vektor, maka nilai *Cosine Similarity* akan semakin mendekati 1 yang menunjukkan tingkat kemiripan yang tinggi. Apabila sudut yang terbentuk semakin besar maka nilai *Cosine Similarity* akan mendekati 0 yang menunjukkan tingkat kemiripan yang rendah atau tidak memiliki kesamaan.

Cosine Similarity digunakan sebagai metrik untuk mengukur tingkat kecocokan antara judul sebagai Q dan transkrip sebagai D karena dapat secara efektif menghitung perbedaan orientasi antar vektor *embedding* pada model *Sentence BERT* dalam ruang semantik tanpa bergantung pada panjang teks (Pitriani, Maylawati, and Gerhana 2025). Vektor tersebut diperoleh melalui proses *mean pooling* yang menghasilkan representasi semantik dari masing-masing teks. Nilai *Cosine Similarity* yang dihasilkan berada pada rentang -1 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan bahwa transkrip memiliki kemiripan yang tinggi dengan judul video, sedangkan nilai yang mendekati 0 menunjukkan bahwa transkrip memiliki kemiripan yang rendah terhadap judul video. Persamaan yang digunakan untuk menghitung nilai *Cosine Similarity* ditunjukkan pada Persamaan 3.2.

$$Sim(Q.D) = \cos(\theta) = \frac{Q \cdot D}{||Q|| ||D||} \quad (3.2)$$

Keterangan:

Q : Vektor judul

D : Vektor transkrip

$Q \cdot D$: *Dot product* (perkalian titik) antara vektor Judul dan Transkrip

$||Q||$: Panjang (magnitudo) vektor Q

$||D||$: Panjang (magnitudo) vektor D

Berikut contoh perhitungan *Cosine Similarity* yang diterapkan pada sebagian dimensi dari vektor embedding hasil mean pooling untuk mempermudah proses perhitungan. Vektor judul ($Q1$) dan transkrip ($D1$) direpresentasikan sebagai berikut:

$$\begin{aligned} Q1 &= [-0.36716, 0.07133, -0.13159, 0.43961, \dots] \\ D1 &= [0.051342, 0.148609, 0.056688, -0.04716, \dots] \end{aligned} \quad (3.3)$$

Perhitungan *Cosine Similarity* dilakukan dengan menghitung nilai *dot product* antara kedua vektor, kemudian dibagi dengan hasil perkalian panjang masing-masing vektor. Perkalian titik antara vektor $Q1$ dan $D1$ dihitung sebagai berikut:

$$\begin{aligned} Q1.D1 &= (-0.36716 \times 0.051342) + (0.07133 \times 0.148609) \\ &+ (-0.13159 \times 0.056688) + (0.43961 \times -0.04716) + \dots \end{aligned} \quad (3.4)$$

Panjang masing-masing vektor dihitung menggunakan akar kuadrat dari jumlah kuadrat setiap dimensinya.

$$\begin{aligned}
 &||Q1| \\
 &= \sqrt{(-0.36716)^2 + (0.07133)^2 + (-0.13159)^2 + (0.43961)^2 + \dots} \\
 &||D1|| \\
 &= \sqrt{(0.051342)^2 + (0.148609)^2 + (-0.056688)^2 + (-0.04716)^2 + \dots}
 \end{aligned} \tag{3.5}$$

Nilai *Cosine Similarity* diperoleh dengan membagi hasil perkalian titik (*dot product*) dengan hasil perkalian panjang kedua vektor tersebut. Berdasarkan perhitungan menggunakan seluruh dimensi embedding (384 dimensi), diperoleh nilai *Cosine Similarity* sebagai berikut:

$$\cos(\theta) = 0.48955 \tag{3.6}$$

Nilai *Cosine Similarity* sebesar 0.48955 menunjukkan tingkat kemiripan antara judul dan transkrip video berdasarkan representasi vektor yang dihasilkan. Nilai tersebut menggambarkan sejauh mana kedekatan makna antara kedua teks yang dibandingkan, dimana semakin tinggi nilai yang diperoleh, maka semakin besar tingkat kemiripan semantik antara judul dan transkrip.

Selanjutnya dilakukan contoh perhitungan *Cosine Similarity* pada data Q2 dan D2 dengan menggunakan sebagian dimensi dari vektor embedding hasil mean

pooling sebagai ilustrasi. Vektor judul (Q2) dan transkrip (D2) direpresentasikan sebagai berikut:

$$\begin{aligned} Q2 &= [-0.12524, 0.216835, -0.05115, -0.06025, \dots] \\ D2 &= [0.050575, -0.05165, 0.018391, -0.03466, \dots] \end{aligned} \quad (3.7)$$

Kemudian dibagi dengan hasil perkalian panjang masing-masing vektor. Perkalian titik antara vektor $Q2$ dan $D2$ dihitung sebagai berikut:

$$\begin{aligned} Q2.D2 &= (-0.12524 \times 0.050575) + (0.216835 \times -0.05165) \\ &+ (-0.05115 \times 0.018391) + (-0.06025 \times -0.03466) + \dots \end{aligned} \quad (3.8)$$

Panjang vektor diperoleh dengan menghitung akar kuadrat dari penjumlahan kuadrat seluruh nilai pada setiap dimensinya.

$$\begin{aligned} ||Q2|| &= \sqrt{(-0.12524)^2 + (0.216835)^2 + (-0.05115)^2 + (-0.06025)^2 + \dots} \\ ||D2|| &= \sqrt{(0.050575)^2 + (-0.05165)^2 + (0.018391)^2 + (-0.03466)^2 + \dots} \end{aligned} \quad (3.9)$$

Nilai *Cosine Similarity* dihitung dengan membagi hasil perkalian titik (*dot product*) antara kedua vektor dengan hasil perkalian panjang masing-masing vektor. Berdasarkan perhitungan menggunakan seluruh dimensi embedding (384 dimensi), diperoleh nilai *Cosine Similarity* sebagai berikut:

$$\cos(\theta) = 0.49772 \quad (3.10)$$

Hasil perhitungan *Cosine Similarity* sebesar 0.49772 menunjukkan tingkat kesamaan antara judul dan transkrip video berdasarkan representasi vektor yang dihasilkan. Nilai tersebut merepresentasikan kedekatan semantik antara kedua teks yang dibandingkan.

3.7 Contoh Perhitungan Lengkap

Pada bagian ini ditampilkan contoh perhitungan untuk menentukan tingkat relevansi antara judul dan transkrip video menggunakan metode *Sentence BERT* dan *Cosine Similarity*. Contoh ini bertujuan untuk memberikan gambaran alur proses perhitungan mulai dari pembentukan representasi vektor hingga penentuan label relevansi. Digunakan dua pasangan data yaitu Q1-D1 dan Q2-D2 sebagai data sampel. Representasi vektor diperoleh dari hasil proses *mean pooling* pada embedding *Sentence BERT* yang menghasilkan vektor berdimensi 384. Pada bagian ini hanya ditampilkan sebagian dimensi sebagai representasi untuk mempermudah pemahaman.

Berdasarkan hasil perhitungan *Cosine Similarity* yang telah dilakukan sebelumnya, diperoleh nilai *similarity* untuk masing-masing pasangan data, yaitu Q1-D1 sebesar 0,48955 dan Q2-D2 sebesar 0,49772. Nilai tersebut menunjukkan tingkat kemiripan semantik antara judul dan transkrip pada masing-masing data.

Selanjutnya dilakukan penentuan nilai *threshold* yang digunakan sebagai batas dalam mengklasifikasikan data ke dalam kategori relevan atau tidak relevan. Dalam penelitian ini digunakan beberapa skenario nilai *threshold*, yaitu 0,10, 0,15,

0,20, dan 0,25 (Holis et al. 2025). Pemilihan nilai-nilai tersebut mengacu pada penelitian terdahulu yang menggunakan *Cosine Similarity* sebagai dasar pengukuran kemiripan teks, dimana nilai *threshold* rendah sering digunakan sebagai batas awal dalam menentukan tingkat relevansi. Penggunaan beberapa variasi nilai ini bertujuan untuk menganalisis pengaruh perubahan nilai *threshold* terhadap hasil klasifikasi yang dihasilkan oleh sistem.

Pendekatan tambahan berupa perhitungan nilai *threshold* berdasarkan rata-rata (*mean*) nilai *similarity* dari data sampel. Perhitungan tersebut dilakukan sebagai berikut:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.11)$$

Keterangan:

$\bar{x}$: nilai rata-rata (*mean*)

n : Jumlah data

x_i : Nilai *similarity* ke- i

Persamaan tersebut kemudian diaplikasikan pada hasil *similarity* Q1-D1 dan Q2-D2 dengan hasil sebagai berikut.

$$\bar{x} = \frac{0.48955 + 0.49772}{2} = 0.493635 \quad (3.12)$$

Nilai *threshold* berbasis rata-rata ini digunakan untuk memperoleh batas yang lebih adaptif terhadap karakteristik data yang digunakan. Dengan demikian, penelitian ini tidak hanya mengandalkan nilai dari penelitian terdahulu, tetapi juga mempertimbangkan distribusi nilai *similarity* pada data yang digunakan.

Berdasarkan nilai *threshold* yang telah ditentukan, maka aturan klasifikasi yang digunakan adalah jika nilai *similarity* lebih besar atau sama dengan *threshold* maka data dikategorikan sebagai relevan (1), sedangkan jika nilai *similarity* lebih kecil dari *threshold* maka data dikategorikan sebagai tidak relevan (0). Pada contoh ini, pasangan data Q1-D1 memiliki nilai *similarity* sebesar 0,48955 yang lebih kecil dari nilai *threshold* rata-rata, sehingga dikategorikan sebagai tidak relevan (0). Sedangkan pasangan data Q2-D2 memiliki nilai *similarity* sebesar 0,49772 yang lebih besar dari *threshold*, sehingga dikategorikan sebagai relevan (1).

Hasil ini menunjukkan bahwa metode *Sentence BERT* yang dikombinasikan dengan *Cosine Similarity* mampu membedakan tingkat relevansi antara judul dan transkrip berdasarkan kedekatan representasi semantik yang dihasilkan. Perbedaan nilai *similarity* yang diperoleh mencerminkan tingkat kesesuaian makna antara kedua pasangan data yang digunakan sebagai contoh.

3.8 Rancangan Pengujian

Untuk mengetahui pengaruh nilai *threshold* terhadap hasil klasifikasi relevansi, penelitian ini menggunakan beberapa skenario pengujian dengan nilai *threshold* yang berbeda. Nilai *threshold* berfungsi sebagai batas keputusan dalam menentukan apakah pasangan judul dan transkrip dikategorikan sebagai relevan atau tidak relevan berdasarkan nilai *Cosine Similarity* yang diperoleh. Semakin

tinggi nilai *threshold*, semakin ketat proses klasifikasi yang dilakukan oleh sistem karena hanya pasangan data dengan tingkat kemiripan yang lebih tinggi yang akan dikategorikan sebagai relevan.

Pada penelitian ini digunakan empat nilai *threshold* yang mengacu pada penelitian terdahulu, yaitu 0.10, 0.15, 0.20, dan 0.25. Selain itu digunakan pula *threshold mean* yang diperoleh dari nilai rata-rata seluruh skor *similarity* pada dataset, yaitu sebesar 0,34. Penggunaan beberapa skenario *threshold* ini bertujuan untuk membandingkan hasil klasifikasi yang dihasilkan serta menentukan nilai *threshold* yang memberikan performa terbaik berdasarkan hasil evaluasi model.

Skenario pengujian yang digunakan dalam penelitian ini dapat dilihat pada Tabel 3.15.

Tabel 3. 15 Skenario Pengujian dengan *Threshold*

No	Skenario	<i>Threshold</i>
1	Skenario 1	0.10
2	Skenario 2	0.15
3	Skenario 3	0.20
4	Skenario 4	0.25
5	Skenario 5	Nilai <i>Mean</i>

Berdasarkan Tabel 3.15, setiap skenario pengujian menggunakan nilai *threshold* yang berbeda untuk melihat perubahan hasil klasifikasi relevansi yang dihasilkan sistem. Seluruh skenario akan diuji menggunakan dataset yang sama sehingga perbedaan hasil yang diperoleh sepenuhnya dipengaruhi oleh perubahan nilai *threshold*. Hasil dari masing-masing skenario kemudian dievaluasi menggunakan *Confusion Matrix*, *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk menentukan nilai *threshold* yang paling optimal.

3.9 Tahap Evaluasi Model

Tahap evaluasi model dilakukan untuk mengukur kinerja sistem dalam mengklasifikasikan pasangan judul dan transkrip ke dalam kategori relevan dan tidak relevan. Evaluasi dilakukan setelah proses perhitungan *Sentence BERT*, *Cosine Similarity*, dan klasifikasi menggunakan berbagai skenario *threshold* selesai dilaksanakan. Hasil klasifikasi yang dihasilkan sistem kemudian dibandingkan dengan label aktual yang telah diperoleh melalui proses labelling. Perbandingan tersebut digunakan untuk membentuk *Confusion Matrix* yang terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

Berdasarkan nilai yang diperoleh dari *Confusion Matrix*, dilakukan perhitungan metrik evaluasi berupa *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk mengukur performa sistem pada setiap skenario *threshold*. Hasil evaluasi dari seluruh skenario pengujian kemudian dibandingkan untuk mengetahui pengaruh perubahan nilai *threshold* terhadap kinerja klasifikasi serta menentukan nilai *threshold* yang menghasilkan performa terbaik. Penjelasan mengenai konsep dan rumus metrik evaluasi yang digunakan dalam penelitian ini telah dijelaskan pada Subbab 2.6.

BAB IV

HASIL DAN PEMBAHASAN

Pada bagian ini merupakan hasil pengolahan data serta pembahasan terhadap kinerja sistem yang telah dibangun. Proses pengujian dilakukan terhadap dataset yang telah melalui tahap pembersihan data sebagaimana dijelaskan pada bab sebelumnya. Hasil yang diperoleh meliputi proses labelisasi data, perhitungan tingkat kemiripan teks menggunakan model *Sentence BERT* dan evaluasi kinerja sistem berdasarkan beberapa metode atau metrik evaluasi. Pada bagian ini juga dilakukan analisis terhadap hasil pengujian yang telah diperoleh untuk mengetahui tingkat keberhasilan sistem dalam mengklasifikasikan relevansi antara judul dan transkrip video. Analisis dilakukan dengan mempertimbangkan pengaruh variasi *threshold* terhadap hasil klasifikasi, sehingga dapat diperoleh nilai *threshold* yang paling optimal. Terdapat juga integrasi hasil penelitian dengan perspektif islam sebagai bentuk penguatan nilai dalam penelitian ini.

4.1 Persiapan Dataset

Pada Dataset yang digunakan dalam penelitian ini merupakan data teks yang terdiri dari pasangan judul dan transkrip video. Data diperoleh dari *Indonesian Stand-Up Comedy Transcription Dataset* (Sumber: Mendeley Data 2020). Dataset ini digunakan sebagai objek penelitian untuk mengukur tingkat kesesuaian antara judul dan isi transkrip menggunakan metode *Sentence BERT* dan *Cosine Similarity*.

Dataset yang digunakan berjumlah sebanyak 2774 data, dimana setiap data memiliki dua atribut utama, yaitu judul dan transkrip. Judul berfungsi sebagai

representasi topik utama dari suatu video, sedangkan transkrip berisi keseluruhan isi pembahasan dalam bentuk teks. Kedua atribut ini menjadi input utama dalam proses pengolahan data pada sistem yang dikembangkan.

Sebelum dilakukan tahap preprocessing dan perhitungan, dataset terlebih dahulu disusun dalam format terstruktur agar dapat diproses oleh sistem. Data disimpan dalam bentuk file spreadsheet (Excel) dengan kolom judul dan transkrip. Contoh data yang digunakan dalam penelitian ini ditunjukkan pada Tabel 4.1.

Tabel 4. 1 Contoh Dataset

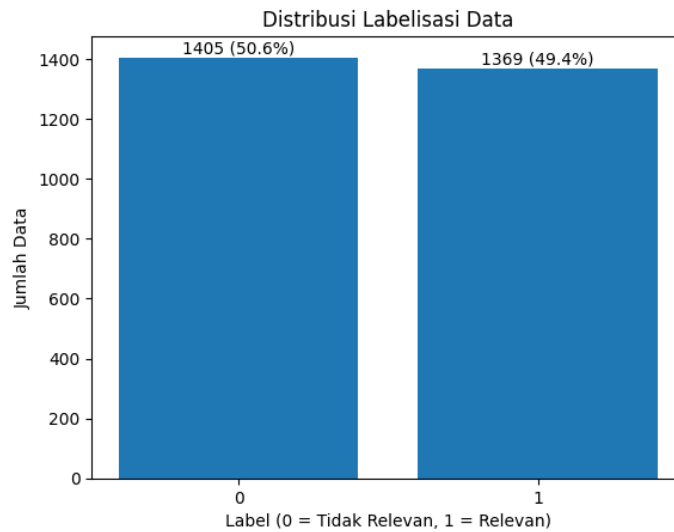
No	Judul	Transkrip
1.	Boah Pengen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu
.....		
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok

Dataset yang digunakan dalam penelitian ini belum memiliki label awal, sehingga belum dapat langsung digunakan untuk proses klasifikasi. Pada tahap selanjutnya akan dilakukan proses pelabelan data menggunakan metode pseudo labeling berdasarkan nilai *Cosine Similarity* antara judul dan transkrip. Dataset ini akan membuat sistem dapat melakukan proses pengolahan mulai dari preprocessing, pembentukan representasi vektor menggunakan *Sentence BERT*, hingga perhitungan tingkat kemiripan antar teks untuk menentukan relevansi antara judul dan transkrip.

4.2 Hasil Labelisasi Data

Proses pelabelan data yang telah dilakukan menghasilkan dataset berlabel yang digunakan sebagai ground truth pada tahap pengujian metode *Sentence BERT* dan *Cosine Similarity*. Hasil pelabelan terhadap 2774 data menunjukkan bahwa sebanyak 1405 data (50,6%) termasuk dalam kategori tidak relevan (label 0), sedangkan 1369 data (49,4%) termasuk dalam kategori relevan (label 1). Distribusi tersebut menunjukkan bahwa jumlah data pada kedua kategori relatif seimbang sehingga dapat digunakan untuk proses evaluasi kinerja sistem.

Untuk memperjelas distribusi jumlah data pada masing-masing kategori label, visualisasi hasil labelisasi disajikan pada Gambar 4.1 berikut.



Gambar 4. 1 Distribusi Lebalisasi Data

Berdasarkan Gambar 4.1, distribusi label menunjukkan proporsi data tidak relevan sedikit lebih besar dibandingkan data relevan. Perbedaan jumlah kedua kategori relatif kecil sehingga tidak terjadi ketimpangan kelas yang signifikan. Kondisi ini memberikan keuntungan pada tahap evaluasi karena model diuji menggunakan data yang memiliki representasi kedua kelas yang hampir seimbang.

Dilakukan validasi manual terhadap 200 data yang dipilih secara acak untuk memastikan kualitas hasil pelabelan otomatis. Hasil validasi menunjukkan bahwa sebanyak 161 data memiliki label yang sesuai dengan hasil pseudo labeling, sedangkan 39 data tidak sesuai. Berdasarkan hasil tersebut diperoleh tingkat kesesuaian sebesar 80,5%, yang menunjukkan bahwa mayoritas label yang dihasilkan secara otomatis telah sesuai dengan penilaian manusia. Maka hasil labelisasi dinilai cukup representatif untuk digunakan sebagai ground truth pada proses evaluasi metode *Sentence BERT* dan *Cosine Similarity*.

4.3 Hasil *Preprocessing*

Bagian ini membahas hasil dari tahap *preprocessing* yang telah dilakukan terhadap data penelitian. Tahapan ini bertujuan untuk membersihkan dan menormalkan teks agar dapat diproses secara lebih optimal pada tahap perhitungan menggunakan metode *Sentence BERT*. Proses yang dilakukan meliputi *cleaning* dan *case folding*, yaitu penghapusan karakter yang tidak diperlukan serta penyeragaman bentuk teks. Proses ini membuat kualitas representasi data menjadi lebih baik sebelum memasuki tahap *embedding*. Hasil dari setiap tahapan akan ditampilkan dalam bentuk perbandingan antara data sebelum dan sesudah diproses.

4.3.1 Hasil *Cleaning*

Pada tahap *cleaning* dilakukan proses pembersihan teks terhadap data judul dengan tujuan untuk menghilangkan karakter yang tidak diperlukan, seperti tanda baca, simbol, angka, serta karakter khusus lainnya yang tidak memiliki kontribusi terhadap makna teks. Proses ini juga mencakup penghapusan elemen tambahan seperti tanda pemisah dan karakter non alfabet yang berpotensi menjadi *noise* dalam data. Proses *cleaning* diharapkan teks judul menjadi lebih sederhana dan terstruktur, sehingga dapat meningkatkan kualitas data sebelum diproses pada tahap selanjutnya. Untuk memperlihatkan perubahan yang terjadi, ditampilkan perbandingan antara data judul sebelum dan sesudah dilakukan proses *cleaning* pada beberapa contoh dataset, sebagaimana disajikan pada Tabel 4.2 berikut.

Tabel 4. 2 Hasil Cleaning Judul pada Contoh Dataset

No	Judul	Setelah Cleaning
1.	Boah Pngen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Boah Pngen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	STAND UP RIO DUMATUBUN GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND
.....		
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	Agungprov Nyamar Jadi Sapu Masjid SUCI OPEN MIC

Bagian Berdasarkan Tabel 4.2, dapat dilihat bahwa proses *cleaning* pada data judul berhasil menghilangkan berbagai karakter yang tidak diperlukan, seperti tanda baca dan simbol pemisah. Teks yang sebelumnya mengandung karakter tambahan menjadi lebih bersih dan hanya menyisakan kata-kata utama yang memiliki makna. Sstruktur judul menjadi lebih rapi dan mudah dipahami tanpa mengubah konteks atau isi informasi yang terkandung di dalamnya. Hal ini menunjukkan bahwa tahap *cleaning* efektif dalam mengurangi *noise* pada data teks sehingga dapat meningkatkan kualitas input pada tahap pemrosesan berikutnya.

Selain diterapkan pada data judul, proses *cleaning* juga dilakukan pada data transkrip yang umumnya memiliki panjang teks yang lebih besar dan kompleks. Karakter yang tidak diperlukan seperti tanda baca, simbol, serta elemen tambahan lainnya dihilangkan untuk menghasilkan teks yang lebih bersih dan terstruktur.

Mengingat transkrip memiliki variasi penulisan yang lebih beragam dibandingkan judul, proses *cleaning* menjadi sangat penting untuk memastikan bahwa data yang digunakan telah terbebas dari *noise* yang dapat mempengaruhi hasil analisis. Perbandingan hasil *cleaning* pada data transkrip ditampilkan pada Tabel 4.3 berikut.

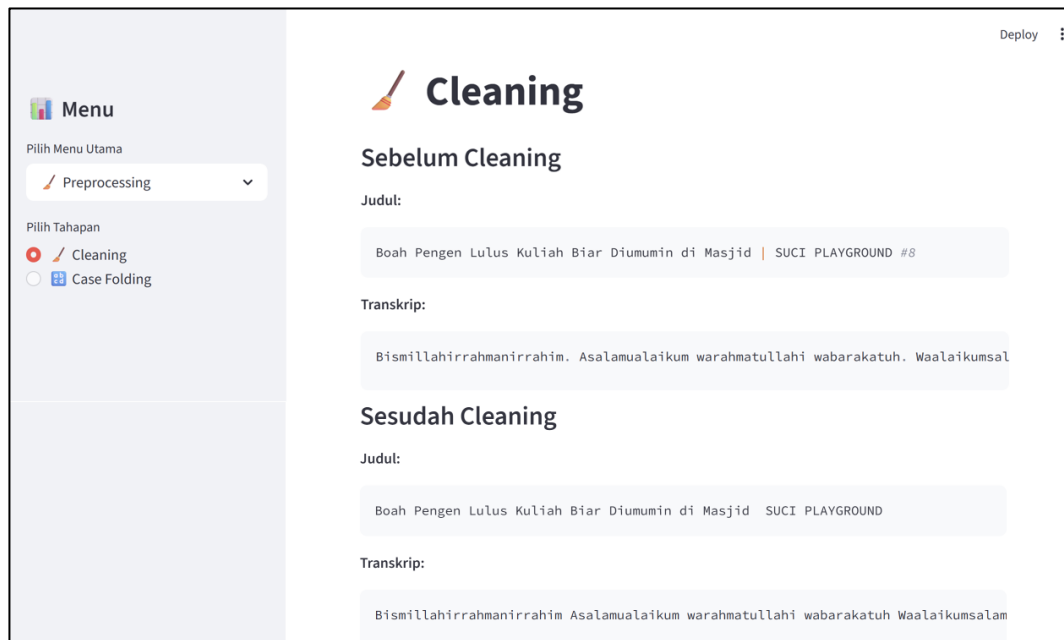
Tabel 4. 3 Hasil Cleaning Transkrip pada Contoh Dataset

No	Transkrip	Setelah Cleaning
1.	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat	Bismillahirrahmanirrahim Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam warahmatullahi wabarakatuh E buat kenal temanteman Bapak Ibu semuanya kenalin buat
2.	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	UBSI mana tepuk tangannya dong buat acara ini Alhamdulillah ya bisa manggung di sini Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong Tepuk tangan buat jujur temanteman Bapakbapak Ibu ya Ee saya ini enggak kuliah gitu enggak pernah wisuda
3.	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati	Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam warahmatullahi wabarakatuh Keren kalian Temanteman Sumpah untuk kali ini gua enggak bakal bohong Jujur dari hati
4.	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya	Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam Kenalin nama saya Rizki Teguh dan saya dari Medan Temanteman di sini ada yang orang Medan Ada Mana Mana orang Medan Kau orang Medan Medan di mana Itu bukan bahasa Medan ya Temanteman ya
5.	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	Asalamualaikum warahmatullahi wabarakatuh BSE Waalaikumsalam warahmatullahi wabarakatuh Luar biasa Gua senang banget bisa stand up di BSI Ini salah satu kampus terbaik di Indonesia Sebenarnya enggak usah tepuk tangan Setiap sanap di kampus gua bilang gitu

.....		
2774.	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok

Berdasarkan Tabel 4.3, terlihat bahwa proses *cleaning* pada data transkrip mampu menghilangkan berbagai karakter yang tidak relevan tanpa mengubah makna utama dari teks. Teks yang sebelumnya mengandung tanda baca dan elemen tambahan menjadi lebih sederhana dan hanya terdiri dari kata-kata yang memiliki nilai informasi. Proses ini juga membantu dalam menyederhanakan struktur kalimat sehingga lebih mudah diproses pada tahap selanjutnya. *Cleaning* pada data transkrip berperan penting dalam meningkatkan kualitas data dan mempersiapkannya untuk proses normalisasi dan representasi vektor.

Untuk memperjelas hasil proses *cleaning* yang telah dilakukan, ditampilkan visualisasi hasil preprocessing melalui aplikasi yang dikembangkan. Tampilan ini menunjukkan perbandingan antara data sebelum dan sesudah proses *cleaning* untuk satu contoh data, baik pada bagian judul maupun transkrip.



Gambar 4. 2 Tampilan Hasil Cleaning pada Aplikasi

Berdasarkan tampilan pada aplikasi dapat dilihat bahwa proses cleaning berhasil menghilangkan karakter yang tidak diperlukan seperti tanda baca dan simbol pemisah tanpa mengubah makna utama dari teks. Pada bagian judul, karakter seperti tanda “|” dan simbol lainnya dihilangkan sehingga menghasilkan teks yang lebih bersih dan terstruktur. Proses cleaning juga mampu menyederhanakan teks dengan menghapus elemen-elemen yang tidak memiliki kontribusi terhadap makna kalimat pada transkrip. Hal ini membuat teks menjadi lebih mudah diproses pada tahap selanjutnya.

4.3.2 Hasil Case Folding

Tahap *case folding* yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil (*lowercase*). Proses ini bertujuan untuk menyeragamkan bentuk kata sehingga tidak terjadi perbedaan representasi antara huruf besar dan huruf kecil yang

sebenarnya memiliki makna yang sama. Tahap ini penting untuk memastikan konsistensi data sebelum dilakukan proses tokenisasi dan embedding pada metode *Sentence BERT*. Perbandingan hasil *case folding* pada data judul ditampilkan pada Tabel 4.4 berikut.

Tabel 4. 4 Hasil Case Folding Judul pada Contoh Dataset

No	Judul	Setelah Case Folding
1.	Boah Pngen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND	boah pngen lulus kuliah biar diumumin di masjid suci playground
2.	STAND UP RIO DUMATUBUN GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND	stand up rio dumatubun gak semua barang cocok dijual di tiktok suci playground
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND	virza disorakin bohong sama penonton suci playground
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND	kehidupan rizky teguh di jakarta tak seindah bayangannya suci playground
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND	cara rahmet mengukur tingkat kebahagiaan mahasiswa suci playground
.....		
2774.	Agungprov Nyamar Jadi Sapu Masjid SUCI OPEN MIC	agungprov nyamar jadi sapu masjid suci open mic

Berdasarkan Tabel 4.4, dapat dilihat bahwa seluruh teks judul telah berhasil dikonversi menjadi huruf kecil secara keseluruhan. Perubahan ini tidak mempengaruhi makna kata, namun mampu menghilangkan perbedaan bentuk penulisan yang disebabkan oleh penggunaan huruf kapital. Proses *case folding* berperan dalam meningkatkan konsistensi data dan memastikan bahwa setiap kata memiliki representasi yang seragam sebelum diproses lebih lanjut.

Proses *case folding* juga diterapkan pada data transkrip untuk memastikan konsistensi penulisan pada seluruh teks. Mengingat transkrip memiliki panjang teks

yang lebih kompleks, penyeragaman huruf menjadi sangat penting agar tidak terjadi perbedaan representasi kata akibat penggunaan huruf besar dan kecil. Hasil dari proses ini ditampilkan dalam bentuk perbandingan antara data sebelum dan sesudah dilakukan *case folding*, sebagaimana disajikan pada Tabel 4.5 berikut.

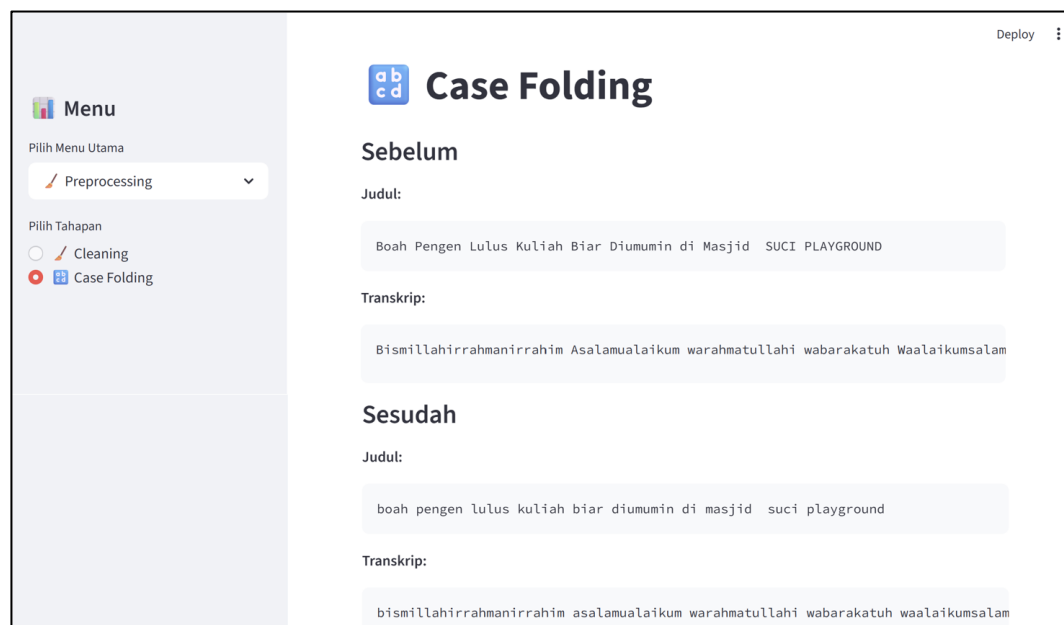
Tabel 4. 5 Hasil Case Folding Transkrip pada Contoh Dataset

No	Transkrip	Setelah Case Folding
1.	Bismillahirrahmanirrahim Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam warahmatullahi wabarakatuh E buat kenal temanteman Bapak Ibu semuanya kenalin buat	bismillahirrahmanirrahim asalamualaikum warahmatullahi wabarakatuh waalaikumsalam warahmatullahi wabarakatuh e buat kenal temanteman bapak ibu semuanya kenalin buat
2.	UBSI mana tepuk tangannya dong buat acara ini Alhamdulillah ya bisa manggung di sini Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong Tepuk tangan buat jujur temanteman Bapakbapak Ibu ya Ee saya ini enggak kuliah gitu enggak pernah wisuda	ubsi mana tepuk tangannya dong buat acara ini alhamdulillah ya bisa manggung di sini temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong tepuk tangan buat jujur temanteman bapakbapak ibu ya ee saya ini enggak kuliah gitu enggak pernah wisuda
3.	Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam warahmatullahi wabarakatuh Keren kalian Temanteman Sumpah untuk kali ini gua enggak bakal bohong Jujur dari hati	asalamualaikum warahmatullahi wabarakatuh waalaikumsalam warahmatullahi wabarakatuh keren kalian temanteman sumpah untuk kali ini gua enggak bakal bohong jujur dari hati
4.	Asalamualaikum warahmatullahi wabarakatuh Waalaikumsalam Kenalin nama saya Rizki Teguh dan saya dari Medan Temanteman di sini ada yang orang Medan Ada Mana Mana orang Medan Kau orang Medan Medan di mana Itu bukan bahasa Medan ya Temanteman ya	asalamualaikum warahmatullahi wabarakatuh waalaikumsalam kenalin nama saya rizki teguh dan saya dari medan temanteman di sini ada yang orang medan ada mana mana orang medan kau orang medan medan di mana itu bukan bahasa medan ya temanteman ya
5.	Asalamualaikum warahmatullahi wabarakatuh BSE Waalaikumsalam warahmatullahi wabarakatuh Luar biasa Gua senang banget bisa stand up di BSI Ini salah satu kampus terbaik di Indonesia Sebenarnya enggak usah tepuk tangan Setiap sanap di kampus gua bilang gitu	asalamualaikum warahmatullahi wabarakatuh bse waalaikumsalam warahmatullahi wabarakatuh luar biasa gua senang banget bisa stand up di bsi ini salah satu kampus terbaik di indonesia sebenarnya enggak usah tepuk tangan setiap sanap di kampus gua bilang gitu
.....		
2774.	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat	saya cepu ya asalamualaikum warahmatullahi wabarakatuh selamat

	malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	malam perkenalkan nama saya agung saya ini polisinya polisi saya provos saya bolos saya tugas di pores depok
--	---	---

Berdasarkan Tabel 4.5, terlihat bahwa seluruh teks transkrip telah berhasil diubah menjadi huruf kecil tanpa mengubah isi dan konteks kalimat. Proses ini menghasilkan teks yang lebih konsisten dan seragam, sehingga dapat meningkatkan kualitas data dalam tahap pemrosesan selanjutnya. Adanya proses *case folding* membuat data menjadi lebih siap untuk digunakan pada tahap tokenisasi dan embedding, karena setiap kata memiliki bentuk representasi yang sama.\

Untuk memberikan gambaran yang lebih jelas mengenai hasil proses case folding, ditampilkan visualisasi hasil perubahan teks melalui aplikasi yang dikembangkan.



Gambar 4. 3 Tampilan Hasil Case Folding pada Aplikasi

Pada tampilan aplikasi terlihat bahwa seluruh teks berhasil diubah menjadi huruf kecil secara konsisten tanpa mengubah susunan kata. Proses ini bertujuan untuk menyeragamkan bentuk teks sehingga perbedaan penulisan huruf besar dan kecil tidak mempengaruhi proses perhitungan pada tahap selanjutnya.

4.4 Hasil dan Analisis Uji Coba

Dilakukan pengujian terhadap sistem pengukuran kemiripan antara judul dan transkrip menggunakan metode *Sentence BERT* dan *Cosine Similarity*. Pengujian dilakukan menggunakan beberapa nilai *threshold* untuk mengetahui pengaruhnya terhadap hasil klasifikasi relevansi data. Proses pengujian dilakukan melalui pengujian satu data pada aplikasi serta pengujian menggunakan seluruh dataset untuk memperoleh hasil klasifikasi secara keseluruhan.

4.4.1 Hasil Perhitungan *Sentence BERT*

Untuk memperjelas proses pembentukan representasi teks menggunakan metode *Sentence BERT*, ditampilkan satu contoh data yang terdiri dari judul dan transkrip melalui antarmuka aplikasi yang telah dibuat.

Gambar 4. 4 Tampilan Input pada Aplikasi

Data yang telah di *input*, terlebih dahulu melalui tahap preprocessing sebagaimana telah dijelaskan pada Subbab 4.3, Selanjutnya data yang telah diproses tersebut digunakan sebagai input pada tahap perhitungan menggunakan model *Sentence BERT*.

Tokenisasi *Sentence BERT* dilakukan menggunakan metode *WordPiece Tokenizer*, di mana satu kata dapat dipecah menjadi beberapa subword. Token khusus seperti [CLS] ditambahkan sebagai representasi keseluruhan kalimat. Adapun *Pseudocode* yang digunakan untuk proses ini disajikan sebagai berikut.

```
text = df['clean_judul'][0]

inputs = tokenizer(
    text,
    return_tensors='pt',
    truncation=True,
    padding=True
)

tokens = tokenizer.convert_ids_to_tokens(
    inputs['input_ids'][0]
)

tokens
```

✓ 0.0s

Gambar 4. 5 *Pseudocode Tokenisasi*

Berdasarkan *pseudocode* pada Gambar 4.5, proses tokenisasi dilakukan terhadap teks hasil *preprocessing* yang tersimpan pada variabel *text*. Teks tersebut diproses menggunakan fungsi *tokenizer* dengan pengaturan *return_tensors='pt'*, *truncation=True*, dan *padding=True* agar sesuai dengan format masukan model *Sentence BERT*. Selanjutnya, token yang dihasilkan dikonversi menjadi bentuk token teks menggunakan fungsi *convert_ids_to_tokens*, sehingga setiap kata dapat direpresentasikan dalam bentuk token maupun *subword*. Hasil tokenisasi disimpan pada variabel *tokens* dan digunakan sebagai masukan pada tahap *token embedding*.

Tampilan hasil tokenisasi menggunakan *Sentence BERT* pada aplikasi dapat dilihat pada Gambar 4.6.

The screenshot shows a web application interface for tokenization. On the left is a sidebar menu with a 'Menu' section containing 'Pilih Menu Utama' (set to 'Sentence BERT') and 'Pilih Tahapan' (with 'Tokenisasi' selected). The main content area is titled 'Tokenisasi' and displays two tables of tokenization results.

	Token Judul
0	[CLS]
1	bo
2	##ah
3	peng
4	##en
5	lulu
6	##s
7	ku
8	##lia
9	##h

	Token Transkrip
0	[CLS]
1	bis
2	##mill
3	##ah
4	##ir
5	##rah
6	##mani
7	##rra
8	##him
9	asa

Gambar 4. 6 Tampilan Hasil Tokenisasi pada Aplikasi

Berdasarkan hasil tokenisasi yang ditampilkan pada Gambar 4.6, terlihat bahwa teks telah berhasil dipecah menjadi token-token penyusun. Kemunculan *subword* yang ditandai dengan simbol ## menunjukkan bahwa kata tertentu dipecah menjadi bagian yang lebih kecil. Hasil tokenisasi ini kemudian digunakan sebagai input pada tahap embedding untuk menghasilkan representasi numerik dari setiap token.

Selanjutnya setiap token yang dihasilkan akan diubah menjadi representasi numerik dalam bentuk vektor melalui proses *token embedding*. Model *Sentence BERT* memetakan setiap token ke dalam vektor berdimensi tertentu yang merepresentasikan makna semantik token tersebut dalam konteks kalimat. Adapun *Pseudocode* yang digunakan untuk proses ini disajikan sebagai berikut.

```
with torch.no_grad():
    outputs = model(**inputs)

    token_embeddings = outputs.last_hidden_state[0]

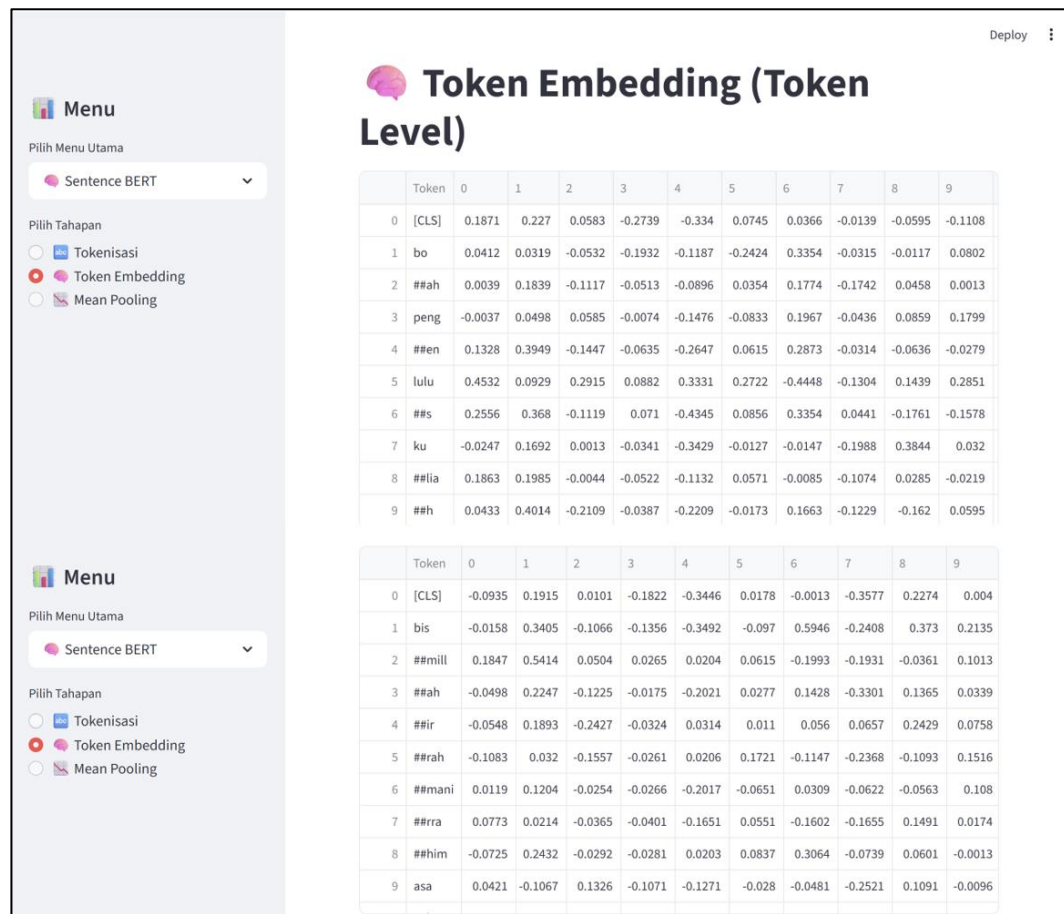
    token_embeddings
```

✓ 0.0s

Gambar 4. 7 *Pseudocode Token Embedding*

Berdasarkan *pseudocode* pada Gambar 4.7, proses *token embedding* dilakukan dengan memasukkan hasil tokenisasi (*inputs*) ke dalam model *Sentence BERT* menggunakan perintah `model(**inputs)`. Proses ini menghasilkan keluaran yang tersimpan pada variabel *outputs*, yang memuat representasi numerik dari setiap token dalam bentuk vektor. Selanjutnya, nilai *token embedding* diambil melalui atribut `last_hidden_state`, kemudian disimpan pada variabel *token_embeddings*. Hasil *token embedding* ini merepresentasikan makna semantik setiap token berdasarkan konteks kalimat dan digunakan sebagai masukan pada tahap *mean pooling* untuk menghasilkan representasi vektor kalimat secara keseluruhan.

Tampilan hasil *token embedding* menggunakan *Sentence BERT* pada aplikasi dapat dilihat pada Gambar 4.8.



Gambar 4. 8 Tampilan Hasil Token Embedding pada Aplikasi

Gambar 4.8 menunjukkan hasil *token embedding*, di mana setiap token direpresentasikan dalam bentuk vektor numerik. Setiap baris pada hasil tersebut menunjukkan satu token beserta nilai vektornya. Nilai-nilai ini mencerminkan informasi semantik yang dipelajari oleh model terhadap token dalam konteks kalimat. Representasi ini selanjutnya digunakan pada tahap mean pooling untuk menghasilkan satu vektor representasi kalimat secara keseluruhan.

Tahap selanjutnya adalah mean pooling dengan menggabungkan seluruh vektor hasil *token embedding* menjadi satu vektor representasi kalimat. Proses ini

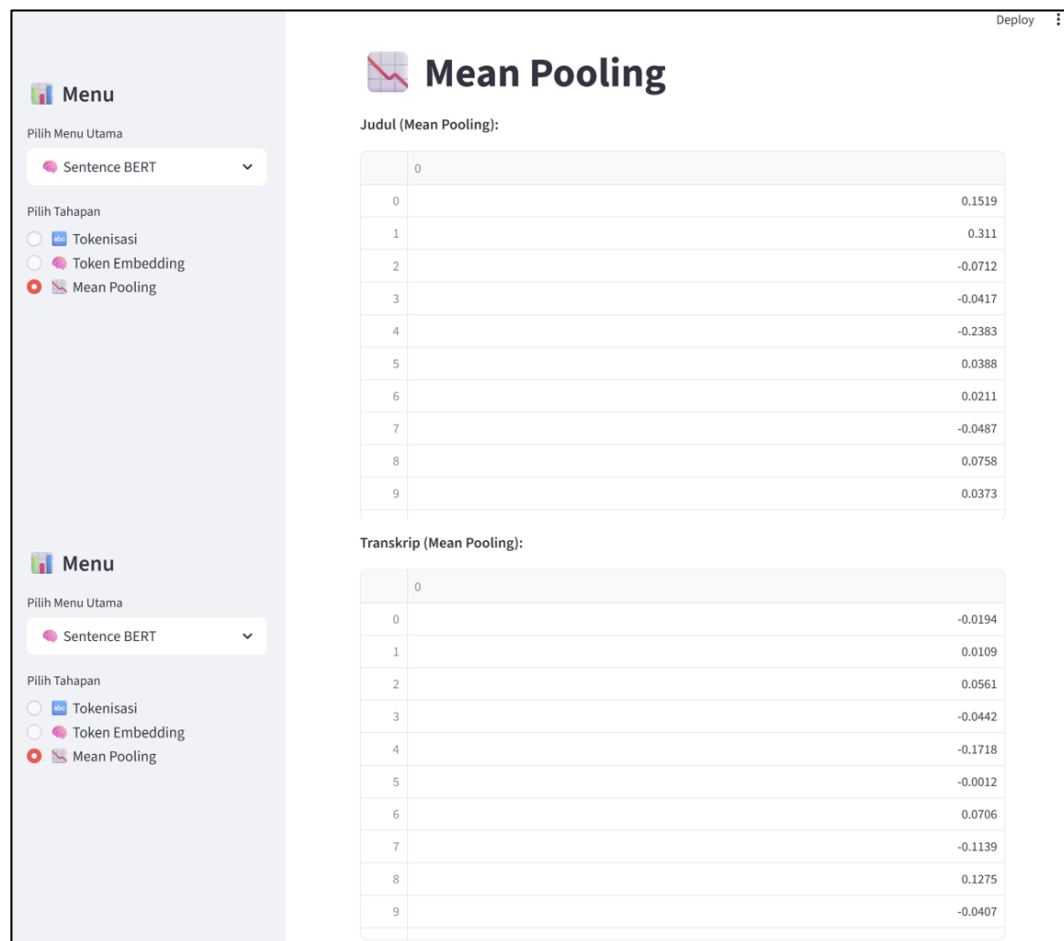
dilakukan dengan menghitung rata-rata dari seluruh vektor token, sehingga diperoleh representasi yang menggambarkan makna keseluruhan kalimat. Adapun *Pseudocode* yang digunakan untuk proses ini disajikan sebagai berikut.

```
mean_embedding = torch.mean(  
    token_embeddings,  
    dim=0  
)  
  
mean_embedding  
✓ 0.0s
```

Gambar 4. 9 *Pseudocode Mean Pooling*

Berdasarkan *pseudocode* pada Gambar 4.9, proses *mean pooling* dilakukan menggunakan fungsi `torch.mean` untuk menghitung nilai rata-rata dari seluruh *token embedding* yang dihasilkan oleh model *Sentence BERT*. Perhitungan dilakukan pada dimensi vektor (`dim=0`) sehingga setiap token dalam kalimat digabungkan menjadi satu vektor representasi. Hasil perhitungan tersebut disimpan pada variabel `mean_embedding` yang merepresentasikan makna keseluruhan kalimat. Vektor hasil *mean pooling* selanjutnya digunakan sebagai masukan pada tahap perhitungan *Cosine Similarity* untuk mengukur tingkat kemiripan antara judul dan transkrip.

Tampilan hasil *mean pooling* menggunakan *Sentence BERT* pada aplikasi dapat dilihat pada Gambar 4.10.



Gambar 4. 10 Tampilan Hasil Mean Pooling pada Aplikasi

Berdasarkan hasil mean pooling pada Gambar 4.10, diperoleh satu vektor representasi untuk setiap teks. Vektor ini merupakan hasil agregasi dari seluruh token embedding yang telah dihasilkan sebelumnya, dan digunakan sebagai representasi semantik kalimat. Vektor tersebut selanjutnya digunakan pada tahap perhitungan kemiripan menggunakan *Cosine Similarity*.

Dimensi vektor yang dihasilkan pada tahap *token embedding* dan *mean pooling* dalam *Sentence BERT* bersifat tetap yaitu sebesar 384. Hal ini dikarenakan model yang digunakan merupakan model pra-latih (*pre-trained model*) dengan

arsitektur tertentu yang telah ditentukan sebelumnya, seperti *all-MiniLM-L6-v2*, yang secara default menghasilkan representasi vektor berdimensi 384. Dimensi ini merupakan ukuran ruang representasi semantik yang digunakan model untuk menangkap makna teks, sehingga setiap token maupun kalimat akan direpresentasikan dalam panjang vektor yang sama agar dapat dibandingkan secara konsisten pada tahap perhitungan kemiripan (Wahyuni et al. 2025).

4.4.2 Hasil Perhitungan *Cosine Similarity*

Setelah diperoleh representasi vektor dari masing-masing teks melalui proses model *Sentence BERT*, langkah selanjutnya adalah menghitung tingkat kemiripan antara vektor judul dan vektor transkrip menggunakan metode *Cosine Similarity*. Adapun *Pseudocode* yang digunakan untuk proses ini disajikan sebagai berikut.

```
similarities = []

for i in range(len(df)):
    sim = cosine_similarity(
        mean_j[i].reshape(1,-1),
        mean_t[i].reshape(1,-1)
    )[0][0]

    similarities.append(sim)

df['similarity_sbert'] = similarities

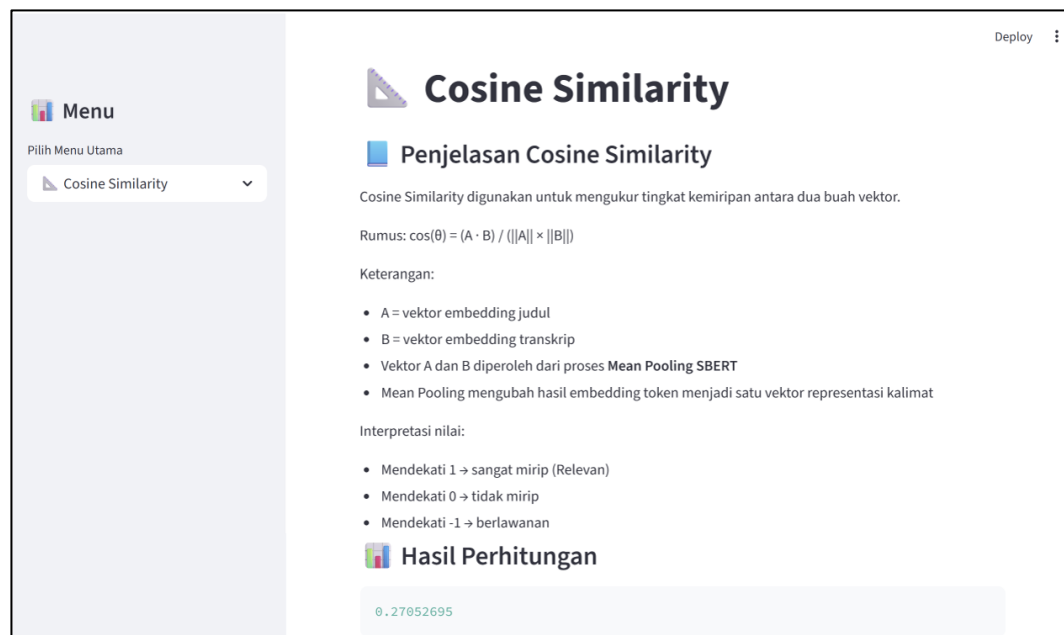
✓ 0.3s
```

Gambar 4. 11 *Pseudocode Cosine Similarity*

Berdasarkan *Pseudocode* pada Gambar 4.11, perhitungan *Cosine Similarity* dilakukan dengan membandingkan vektor hasil *mean pooling* dari judul (*mean_j*) dan transkrip (*mean_t*). Pada setiap data, fungsi *cosine_similarity* digunakan untuk

menghitung tingkat kemiripan antara kedua vektor tersebut, kemudian hasilnya disimpan ke dalam variabel *similarities*. Seluruh nilai kemiripan yang diperoleh selanjutnya disimpan pada kolom *similarity_sbert* dan digunakan sebagai dasar dalam proses klasifikasi relevansi berdasarkan nilai *threshold* yang telah ditentukan. Semakin tinggi nilai *similarity* yang dihasilkan, semakin besar tingkat keterkaitan semantik antara judul dan transkrip yang dibandingkan.

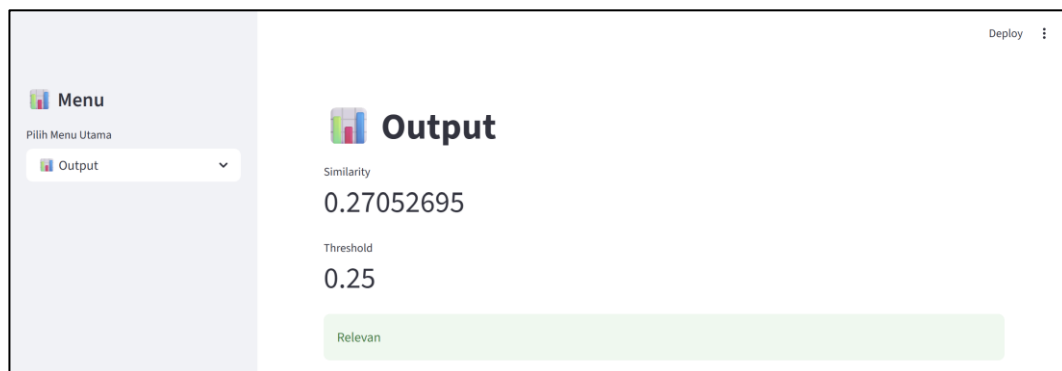
Tampilan hasil *Cosine Similarity* menggunakan *Sentence BERT* pada aplikasi dapat dilihat pada Gambar 4.12.



Gambar 4. 12 Tampilan Hasil *Cosine Similarity* pada Aplikasi

Diperoleh nilai *Cosine Similarity* sebesar 0,2705. Hasil ini kemudian digunakan sebagai dasar dalam proses klasifikasi dengan membandingkan nilai tersebut terhadap nilai *threshold* yang telah ditentukan.

Selanjutnya sistem melakukan klasifikasi relevansi dengan membandingkan nilai *similarity* terhadap *threshold* sebesar 0,25 untuk menentukan *output* apakah data termasuk relevan atau tidak relevan.



Gambar 4. 13 Tampilan Hasil *Output* pada Aplikasi

Berdasarkan Gambar 4.13 diperoleh nilai *similarity* sebesar 0,27052695 dengan nilai *threshold* sebesar 0,25. Karena nilai *similarity* lebih besar daripada *threshold*, sistem mengklasifikasikan pasangan data tersebut sebagai relevan. Hasil ini menunjukkan bahwa proses klasifikasi pada aplikasi dilakukan dengan membandingkan nilai kemiripan terhadap batas *threshold* yang telah ditentukan.

4.4.3 Skenario Pengujian

Pengujian dalam penelitian ini dilakukan menggunakan dataset secara keseluruhan yang berjumlah 2774 data. Pada tahap ini dilakukan pengujian sistem menggunakan beberapa skenario *threshold* untuk mengetahui pengaruh perubahan nilai *threshold* terhadap hasil klasifikasi relevansi data. Dalam penelitian ini digunakan beberapa variasi nilai *threshold* yang sudah di jelaskan pada tabel 3.x yaitu 0,10, 0,15, 0,20, 0,25, serta nilai *threshold* yang diperoleh dari hasil perhitungan rata-rata (*mean*) nilai *similarity*. Nilai-nilai tersebut dipilih berdasarkan

penelitian terdahulu yang menggunakan *Cosine Similarity* sebagai dasar pengukuran kemiripan teks, serta untuk melihat sensitivitas sistem terhadap perubahan nilai ambang batas.

Dilakukan perhitungan nilai rata-rata (*mean*) dari seluruh nilai *similarity* yang dihasilkan pada dataset. Berdasarkan hasil perhitungan terhadap 2774 data, diperoleh nilai sebagai berikut:

$$\bar{x} = \frac{956}{2774} = 0.34467953 \quad (4.1)$$

Nilai ini kemudian digunakan sebagai salah satu skenario dalam pengujian untuk menentukan klasifikasi relevansi data. Untuk memberikan visualisasi mengenai penerapan masing-masing nilai *threshold* pada sistem, ditampilkan hasil eksperimen menggunakan beberapa skenario *threshold*, yaitu 0,10, 0,15, 0,20, 0,25, dan *threshold mean*. Setiap eksperimen menunjukkan hasil klasifikasi data berdasarkan nilai *similarity* yang diperoleh dari proses *Sentence BERT* dan *Cosine Similarity*. Adapun *Pseudocode* yang digunakan untuk proses ini disajikan sebagai berikut.

```
threshold = df['similarity_sbert'].mean() #Menentukan Nilai Threshold

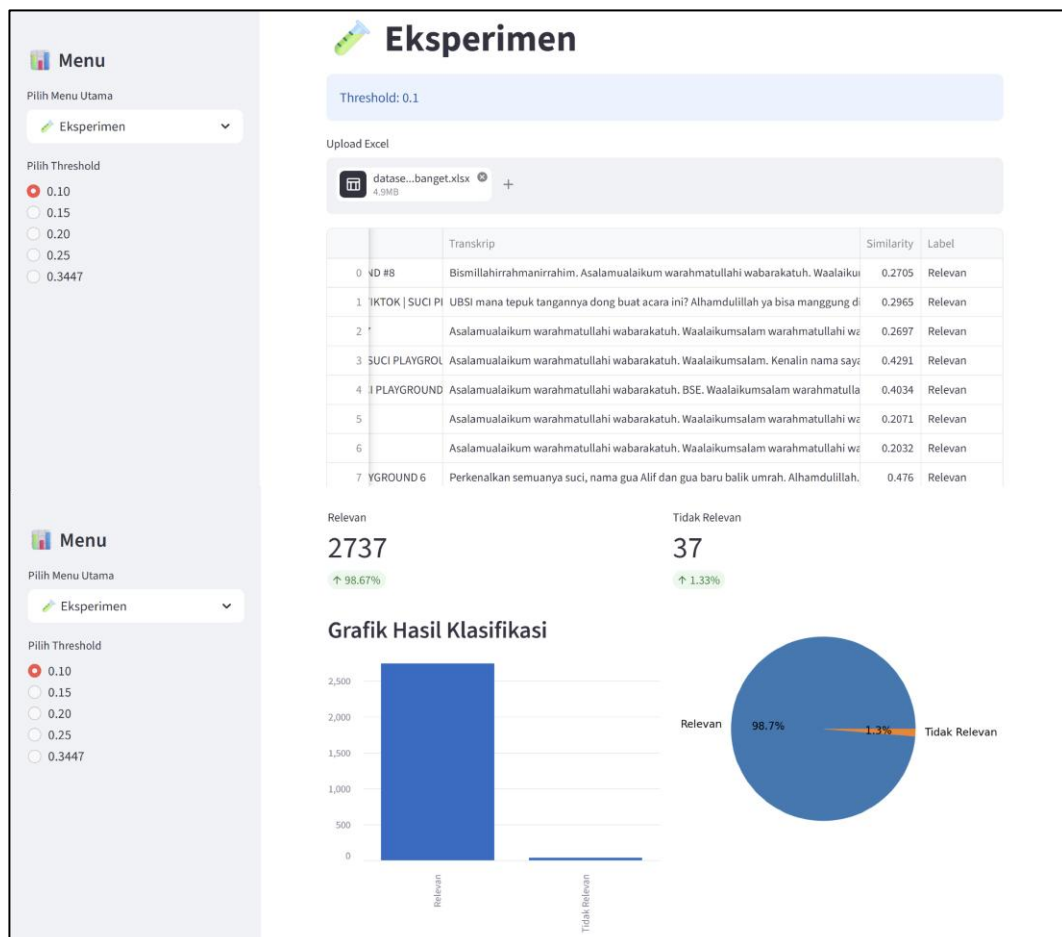
df['label_sbert'] = df['similarity_sbert'].apply(
    lambda x: 1 if x >= threshold else 0 #Bandingkan similarity dengan threshold
)

df[['similarity_sbert', 'label_sbert']] #Menampilkan hasil klasifikasi
✓ 0.0s
```

Gambar 4. 14 *Pseudocode Pemilihan Threshold*

Pseudocode Pada gambar 4.14 menunjukkan bahwa nilai *threshold* ditentukan menggunakan *mean* seluruh nilai *similarity_sbert* yang tersimpan pada dataset. Nilai *mean* tersebut menghasilkan *threshold mean* sebesar 0,34 yang kemudian digunakan sebagai batas klasifikasi. Setiap nilai *similarity* dibandingkan dengan *threshold mean* menggunakan fungsi *lambda*, sehingga diperoleh label 1 untuk data relevan dan label 0 untuk data tidak relevan.

Tampilan hasil masing-masing *threshold* pada aplikasi dapat dilihat pada Gambar 4.15.



Gambar 4. 15 Hasil Eksperimen Skenario Pengujian 0.10 pada Aplikasi

Berdasarkan Gambar 4.15, hasil eksperimen skenario pengujian menggunakan *threshold* 0,10 menunjukkan bahwa sebagian besar data diklasifikasikan sebagai relevan. Dari total 2774 data, sebanyak 2737 data atau sekitar 98,7% masuk ke kategori relevan, sedangkan 37 data atau sekitar 1,3% dikategorikan sebagai tidak relevan. Dominasi data relevan terlihat pada *bar chart* dan *pie chart* yang ditampilkan sistem. Hal ini menunjukkan bahwa penggunaan *threshold* yang rendah membuat sistem cenderung menganggap sebagian besar pasangan judul dan transkrip memiliki keterkaitan semantik.

Berikut merupakan beberapa hasil klasifikasi data menggunakan *threshold* 0,10 yang menampilkan nilai *similarity* dan label relevansi yang dihasilkan sistem.

Tabel 4. 6 Hasil Klasifikasi Data Berdasarkan *Threshold* 0.10

No	Judul	Transkrip	Similarity	Label
1.	Boah Pngen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat	0,270527	1 (Relevan)
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	0,296468	1 (Relevan)
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi	0,269726	1 (Relevan)

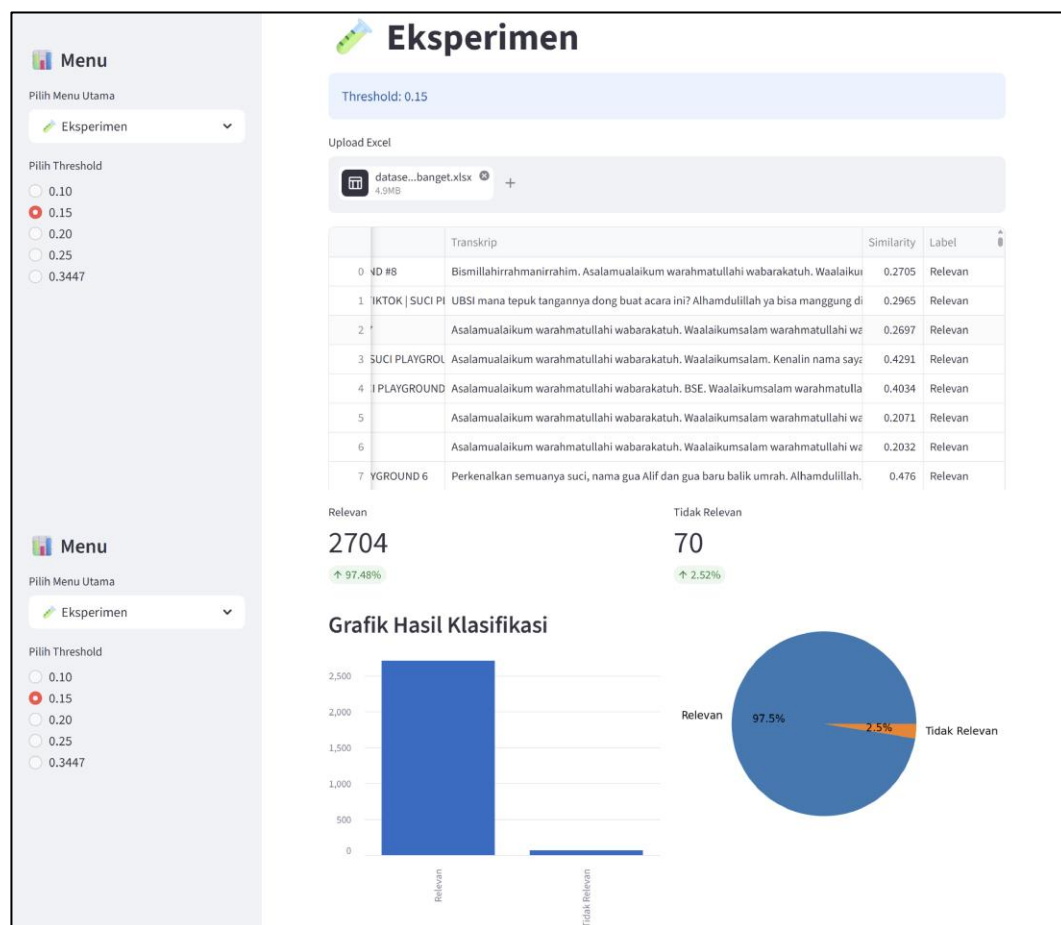
	PLAYGROUND #7	wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati		
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya	0,429099	1 (Relevan)
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	0,403429	1 (Relevan)
.....				
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	0,253936	1 (Relevan)

Berdasarkan Tabel 4.7, seluruh data diklasifikasikan sebagai relevan karena nilai *similarity* yang diperoleh berada di atas *threshold* 0,10. Nilai *similarity* pada contoh data berkisar antara 0,253936 hingga 0,429099, sehingga sistem masih

menganggap pasangan judul dan transkrip memiliki keterkaitan semantik. Meskipun beberapa pasangan data memiliki tingkat kemiripan yang tidak terlalu tinggi, nilai tersebut tetap melewati batas *threshold* yang digunakan sehingga label yang dihasilkan adalah relevan.

Hasil ini menunjukkan bahwa penggunaan *threshold* 0,10 membuat proses klasifikasi menjadi cukup longgar. Sistem cenderung menerima sebagian besar pasangan data sebagai relevan karena batas minimum kemiripan yang digunakan relatif rendah. Kondisi tersebut sejalan dengan hasil distribusi klasifikasi pada aplikasi, di mana jumlah data relevan jauh lebih dominan dibandingkan data tidak relevan.

Selanjutnya dilakukan pengujian menggunakan nilai *threshold* sebesar 0,15 untuk melihat perubahan hasil klasifikasi relevansi pada sistem.



Gambar 4. 16 Hasil Eksperimen Skenario Pengujian 0.15 pada Aplikasi

Berdasarkan Gambar 4.16, hasil eksperimen skenario pengujian menggunakan *threshold* 0,15 menunjukkan bahwa sebagian besar data masih diklasifikasikan sebagai relevan. Dari total 2774 data, sebanyak 2704 data atau sekitar 97,5% masuk ke kategori relevan, sedangkan 70 data atau sekitar 2,5% dikategorikan sebagai tidak relevan. Peningkatan pada *threshold* ini membuat sistem mulai lebih selektif dalam menentukan relevansi pasangan judul dan transkrip video, sehingga beberapa data dengan nilai *similarity* rendah tidak lagi diklasifikasikan sebagai relevan.

Berikut merupakan beberapa hasil klasifikasi data menggunakan *threshold* 0,15 yang menampilkan nilai *similarity* dan label relevansi yang dihasilkan sistem.

Tabel 4. 7 Hasil Klasifikasi Data Berdasarkan *Threshold* 0.15

No	Judul	Transkrip	Similarity	Label
1.	Boah Pengen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat	0,270527	1 (Relevan)
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	0,296468	1 (Relevan)
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati	0,269726	1 (Relevan)
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya	0,429099	1 (Relevan)

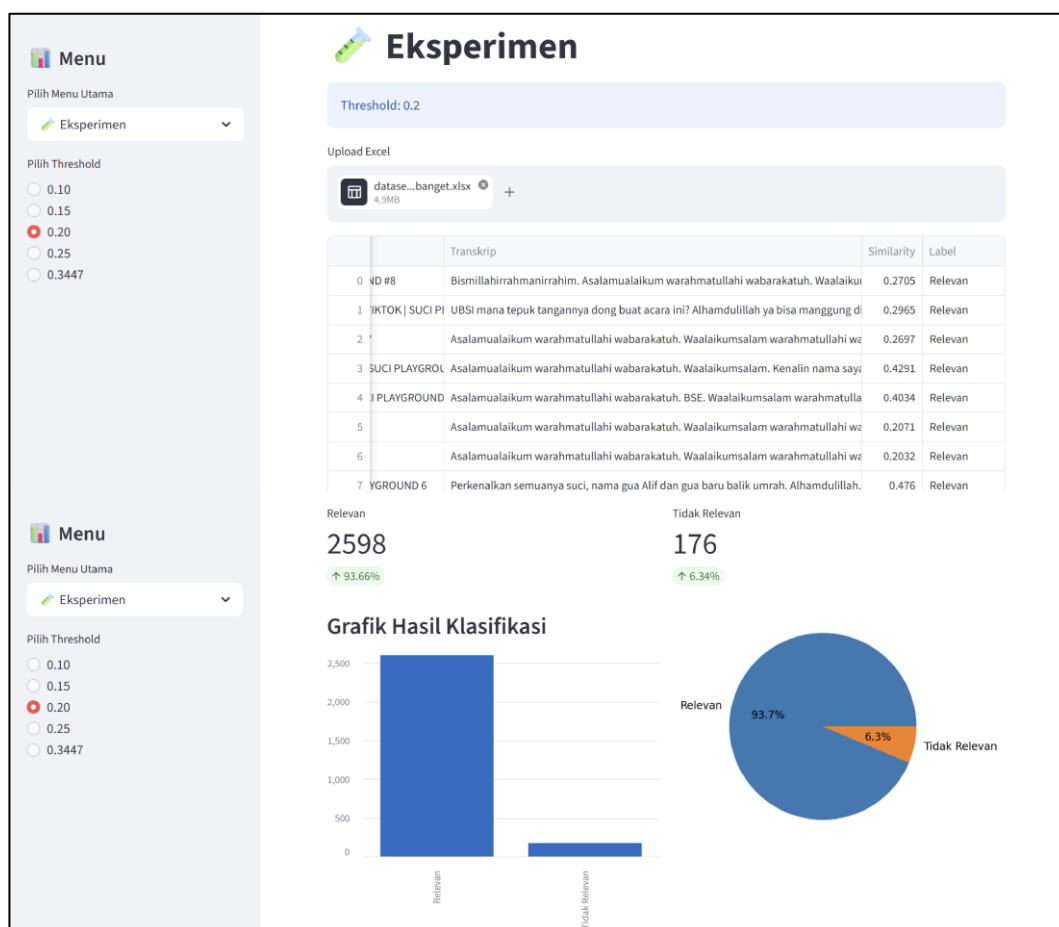
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	0,403429	1 (Relevan)
.....				
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	0,253936	1 (Relevan)

Berdasarkan Tabel 4.8, seluruh data masih diklasifikasikan sebagai relevan karena nilai *similarity* yang diperoleh berada di atas *threshold* 0,15. Nilai *similarity* pada contoh data berada pada rentang 0,253936 hingga 0,429099, sehingga sistem masih menganggap pasangan judul dan transkrip video memiliki keterkaitan semantik. Data dengan nilai *similarity* tinggi seperti 0,429099 dan 0,403429, menunjukkan tingkat kemiripan yang lebih kuat sehingga dengan jelas masuk ke kategori relevan.

Penggunaan *threshold* 0,15 membuat sistem mulai lebih selektif dalam proses klasifikasi. Data dengan nilai *similarity* yang mendekati batas *threshold* memiliki peluang lebih besar untuk dikategorikan sebagai tidak relevan apabila nilai *threshold* ditingkatkan kembali. Hal ini menunjukkan bahwa peningkatan

threshold mempengaruhi proses penyaringan data, sehingga sistem tidak terlalu longgar dalam menentukan relevansi pasangan judul dan transkrip video.

Selanjutnya dilakukan pengujian menggunakan nilai *threshold* sebesar 0,20 untuk melihat perubahan hasil klasifikasi relevansi pada sistem.



Gambar 4. 17 Hasil Eksperimen Skenario Pengujian 0.20 pada Aplikasi

Berdasarkan Gambar 4.17, hasil eksperimen skenario pengujian menggunakan *threshold* 0,20 menunjukkan bahwa sistem mulai melakukan klasifikasi yang lebih selektif dibandingkan *threshold* sebelumnya. Dari total 2774 data, sebanyak 2598 data atau sekitar 93,7% dikategorikan sebagai relevan,

sedangkan 176 data atau sekitar 6,3% masuk ke kategori tidak relevan. Perubahan distribusi data tersebut terlihat pada *bar chart* dan *pie chart* yang menunjukkan peningkatan jumlah data tidak relevan. Kenaikan nilai *threshold* menyebabkan sistem lebih ketat dalam menentukan keterkaitan semantik antara judul dan transkrip video, sehingga data dengan nilai *similarity* rendah mulai dikelompokkan ke dalam kategori tidak relevan.

Berikut merupakan beberapa hasil klasifikasi data menggunakan *threshold* 0,20 yang menampilkan nilai *similarity* dan label relevansi yang dihasilkan sistem.

Tabel 4. 8 Hasil Klasifikasi Data Berdasarkan *Threshold* 0.20

No	Judul	Transkrip	<i>Similarity</i>	Label
1.	Boah Pengen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat	0,270527	1 (Relevan)
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	0,296468	1 (Relevan)
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati	0,269726	1 (Relevan)

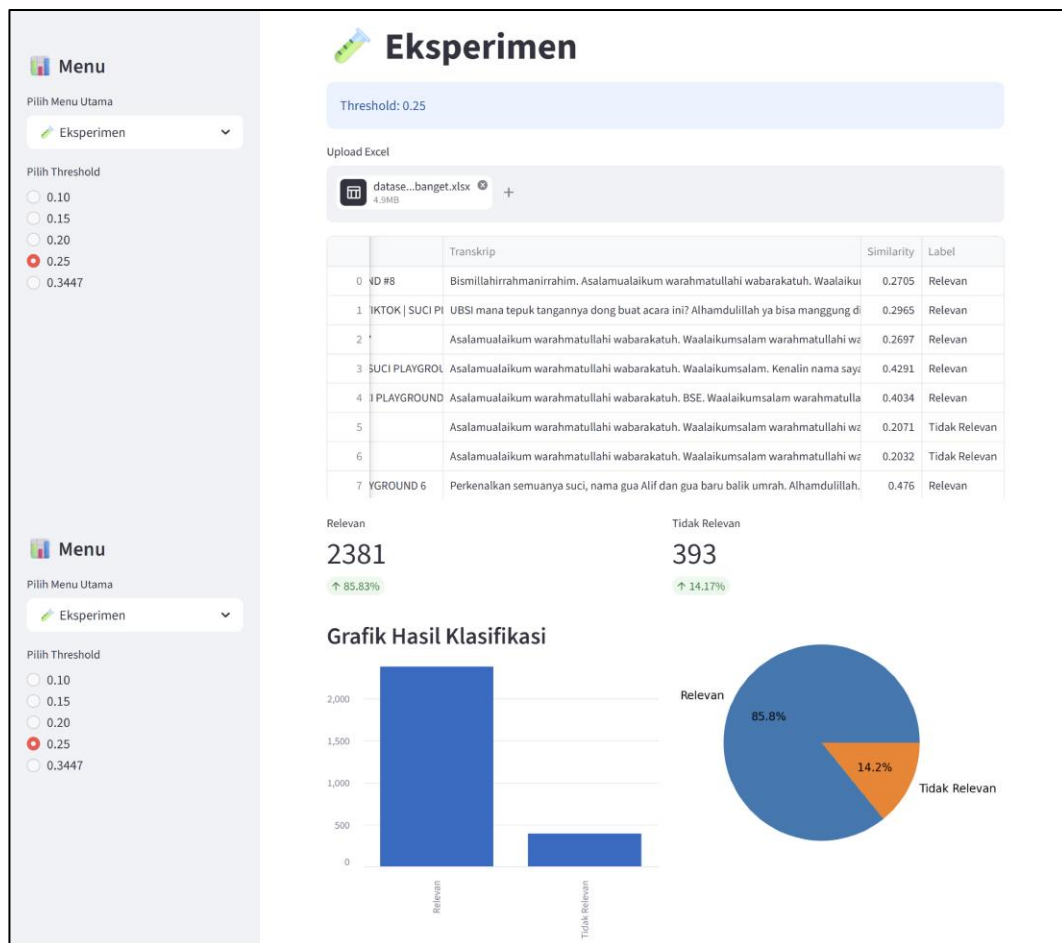
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya	0,429099	1 (Relevan)
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	0,403429	1 (Relevan)
.....				
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	0,253936	1 (Relevan)

Berdasarkan Tabel 4.9, seluruh contoh data masih diklasifikasikan sebagai relevan karena nilai *similarity* yang diperoleh berada di atas *threshold* 0,20. Nilai *similarity* pada data yang di tampilkan berkisar antara 0,253936 hingga 0,429099, sehingga pasangan judul dan transkrip masih dianggap memiliki keterkaitan semantik oleh sistem. Data dengan nilai *similarity* tinggi seperti 0,429099 dan

0,403429, menunjukkan tingkat kemiripan yang kuat sehingga tetap berada pada kategori relevan.

Penggunaan threshold 0,20 membuat proses klasifikasi menjadi lebih selektif. Data dengan nilai similarity yang mendekati batas *threshold* memiliki peluang lebih besar untuk dikategorikan sebagai tidak relevan. Hal ini menunjukkan bahwa peningkatan *threshold* membuat sistem lebih ketat dalam menentukan relevansi antara judul dan transkrip.

Selanjutnya dilakukan pengujian menggunakan nilai *threshold* sebesar 0,25 untuk melihat perubahan hasil klasifikasi relevansi pada sistem.



Gambar 4. 18 Hasil Eksperimen Skenario Pengujian 0.25 pada Aplikasi

Berdasarkan Gambar 4.18, hasil eksperimen scenario pengujian menggunakan *threshold* 0,25 menunjukkan bahwa sistem melakukan klasifikasi yang lebih selektif dibandingkan *threshold* sebelumnya. Dari total 2774 data, sebanyak 2381 data atau sekitar 85,8% dikategorikan sebagai relevan, sedangkan 393 data atau sekitar 14,2% masuk ke kategori tidak relevan. Perubahan distribusi data terlihat cukup jelas pada *bar chart* dan *pie chart*, di mana jumlah data tidak relevan meningkat dibandingkan pengujian sebelumnya. Hal ini menunjukkan bahwa peningkatan nilai *threshold* membuat sistem lebih ketat dalam menentukan keterkaitan semantik antara judul dan transkrip video, sehingga hanya data dengan nilai *similarity* yang cukup tinggi yang tetap diklasifikasikan sebagai relevan.

Berikut merupakan beberapa hasil klasifikasi data menggunakan *threshold* 0,25 yang menampilkan nilai *similarity* dan label relevansi yang dihasilkan sistem.

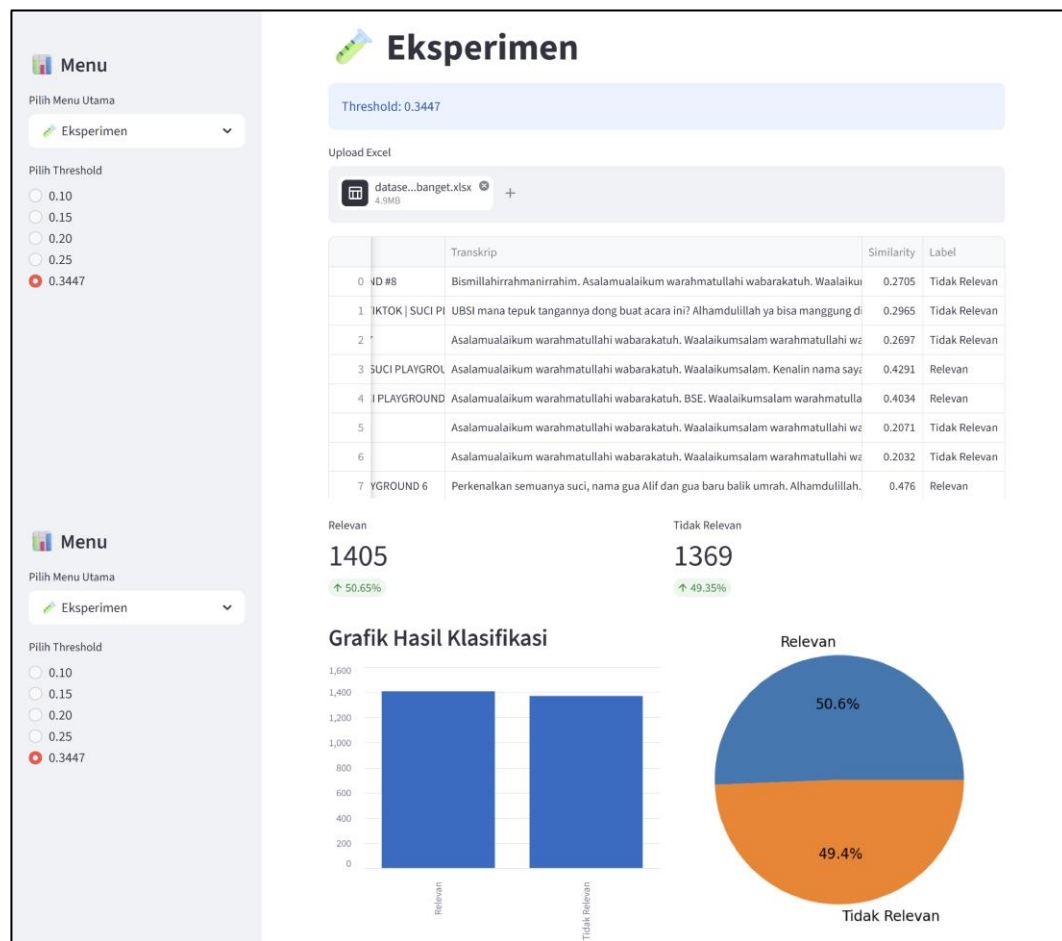
Tabel 4. 9 Hasil Klasifikasi Data Berdasarkan *Threshold* 0.25

No	Judul	Transkrip	<i>Similarity</i>	Label
1.	Boah Pengen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenal in buat	0,270527	1 (Relevan)
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	0,296468	1 (Relevan)

3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati	0,269726	1 (Relevan)
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana? Itu bukan bahasa Medan ya, Temanteman ya	0,429099	1 (Relevan)
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	0,403429	1 (Relevan)
.....				
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	0,253936	1 (Relevan)

Berdasarkan Tabel 4.10, hasil klasifikasi menggunakan *threshold* 0,25 menunjukkan bahwa sistem mulai melakukan penyaringan data secara lebih ketat dibandingkan *threshold* sebelumnya. Data dengan nilai *similarity* yang berada di atas 0,25 tetap diklasifikasikan sebagai relevan, sedangkan data dengan nilai *similarity* di bawah *threshold* akan masuk ke kategori tidak relevan. Pada tabel terlihat bahwa sebagian besar data masih memiliki nilai *similarity* di atas *threshold*, seperti 0,270527, 0,296468, 0,429099, dan 0,403429 sehingga tetap diberi label relevan. Nilai *similarity* yang berada dekat dengan batas *threshold*, seperti 0,253936, menunjukkan bahwa sistem mulai membedakan tingkat keterkaitan semantik antar teks secara lebih selektif. Hasil ini menunjukkan bahwa peningkatan *threshold* menyebabkan proses klasifikasi menjadi lebih ketat karena hanya pasangan judul dan transkrip video dengan tingkat kemiripan yang cukup tinggi yang dipertahankan sebagai relevan. *Threshold* 0,25 menghasilkan distribusi klasifikasi yang lebih seimbang dibandingkan *threshold* 0,10, 0,15, dan 0,20, karena jumlah data tidak relevan mulai meningkat seiring bertambahnya batas minimal *similarity* yang digunakan sistem.

Selanjutnya dilakukan pengujian menggunakan nilai *threshold mean* untuk melihat perubahan hasil klasifikasi relevansi pada sistem.



Gambar 4. 19 Hasil Eksperimen Skenario Pengujian *Threshold Mean*

Berdasarkan Gambar 4.19, hasil pengujian menggunakan *threshold mean* sebesar 0,34 menunjukkan bahwa proses klasifikasi menjadi lebih seimbang dibandingkan skenario *threshold* sebelumnya. Dari total 2774 data, sebanyak 1405 data atau sekitar 50,6% diklasifikasikan sebagai relevan, sedangkan 1369 data atau sekitar 49,4% masuk ke kategori tidak relevan. Hasil tersebut terlihat pada tampilan tabel, *bar chart*, dan *pie chart* yang menunjukkan distribusi data hampir seimbang antara kedua kategori. Penggunaan *threshold mean* menyebabkan sistem menjadi lebih selektif karena hanya pasangan judul dan transkrip video dengan nilai

similarity yang cukup tinggi yang dikategorikan relevan. Kondisi ini menunjukkan bahwa *threshold mean* mampu mengurangi dominasi label relevan yang sebelumnya terjadi pada *threshold* rendah.

Tabel 4. 10 Hasil Klasifikasi Data Berdasarkan *Threshold Mean*

No	Judul	Transkrip	Similarity	Label
1.	Boah Pengen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahmanirrahim. Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. E buat kenal temanteman, Bapak, Ibu semuanya kenalin buat	0,270527	0 (Tidak Relevan)
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI TIKTOK SUCI PLAYGROUND #8	UBSI mana tepuk tangannya dong buat acara ini? Alhamdulillah ya bisa manggung di sini. Temanteman tepuk tangan dong buat tadi yang para wisudawan tuh dong. Tepuk tangan buat jujur temanteman, Bapakbapak, Ibu ya. Ee saya ini enggak kuliah gitu, enggak pernah wisuda	0,296468	0 (Tidak Relevan)
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam warahmatullahi wabarakatuh. Keren kalian, Temanteman. Sumpah untuk kali ini gua enggak bakal bohong. Jujur dari hati	0,269726	0 (Tidak Relevan)
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh dan saya dari Medan. Temanteman di sini ada yang orang Medan? Ada. Mana? Mana orang Medan? Kau orang Medan? Medan di mana?	0,429099	1 (Relevan)

		Itu bukan bahasa Medan ya, Temanteman ya		
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullahi wabarakatuh. BSE. Waalaikumsalam warahmatullahi wabarakatuh. Luar biasa. Gua senang banget bisa stand up di BSI. Ini salah satu kampus terbaik di Indonesia. Sebenarnya enggak usah tepuk tangan. Setiap sanap di kampus gua bilang gitu	0,403429	1 (Relevan)
.....				
2774.	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullahi wabarakatuh selamat malam perkenalkan nama saya Agung saya ini polisinya polisi saya Provos saya bolos saya tugas di pores Depok	0,253936	0 (Tidak Relevan)

Berdasarkan Tabel 4.11, penggunaan *threshold mean* sebesar 0,34 membuat proses klasifikasi menjadi lebih selektif. Data dengan nilai *similarity* di bawah *threshold*, seperti 0,270527, 0,296468, 0,269726, dan 0,253936 dikategorikan sebagai tidak relevan, sedangkan data dengan *similarity* lebih tinggi, seperti 0,429099 dan 0,403429 tetap diklasifikasikan sebagai relevan. Hasil ini menunjukkan bahwa *threshold mean* mampu menghasilkan distribusi klasifikasi yang lebih seimbang karena sistem hanya mempertahankan pasangan judul dan transkrip video dengan tingkat kemiripan semantik yang cukup tinggi sebagai data relevan.

Untuk memberikan gambaran keseluruhan mengenai hasil eksperimen pengujian pada setiap nilai *threshold*, dilakukan rekapitulasi jumlah data relevan

dan tidak relevan yang dihasilkan sistem. Rekapitulasi hasil klasifikasi tersebut ditampilkan pada Tabel 4.12 berikut.

Tabel 4. 11 Hasil Pengujian *Threshold*

<i>Threshold</i>	Relevan (1)	Tidak Relevan (0)	Total
0.10	2737	37	2774
0.15	2704	70	
0.20	2598	176	
0.25	2381	393	
0.34467953	1405	1369	

Berdasarkan Tabel 4.12, terlihat bahwa nilai threshold memiliki pengaruh yang signifikan terhadap hasil klasifikasi relevansi antara judul dan transkrip. Semakin rendah nilai threshold yang digunakan, semakin banyak data yang diklasifikasikan sebagai relevan. Hal ini terlihat pada threshold 0,10 yang menghasilkan 2737 data relevan dan hanya 37 data tidak relevan. Kondisi tersebut menunjukkan bahwa batas kemiripan yang rendah membuat sistem lebih mudah mengategorikan data sebagai relevan.

Seiring meningkatnya nilai threshold menjadi 0,15, 0,20, dan 0,25, jumlah data relevan terus mengalami penurunan, sedangkan jumlah data tidak relevan meningkat. Pada threshold 0,25, jumlah data relevan berkurang menjadi 2381 data dan data tidak relevan meningkat menjadi 393 data. Fenomena ini menunjukkan bahwa sistem menjadi semakin selektif karena hanya data dengan tingkat kemiripan yang lebih tinggi yang dapat dikategorikan sebagai relevan.

Pada threshold mean sebesar 0,34, distribusi klasifikasi menjadi lebih seimbang dengan 1405 data relevan dan 1369 data tidak relevan. Hasil ini menunjukkan bahwa penggunaan threshold yang lebih tinggi mampu memperketat proses klasifikasi sehingga mengurangi kecenderungan sistem untuk memberikan

label relevan secara berlebihan. Dapat disimpulkan bahwa semakin tinggi nilai threshold maka semakin ketat kriteria penentuan relevansi, sedangkan semakin rendah nilai threshold maka semakin longgar proses klasifikasi sehingga jumlah data yang dikategorikan relevan menjadi lebih banyak.

4.5 Evaluasi Kinerja Sistem

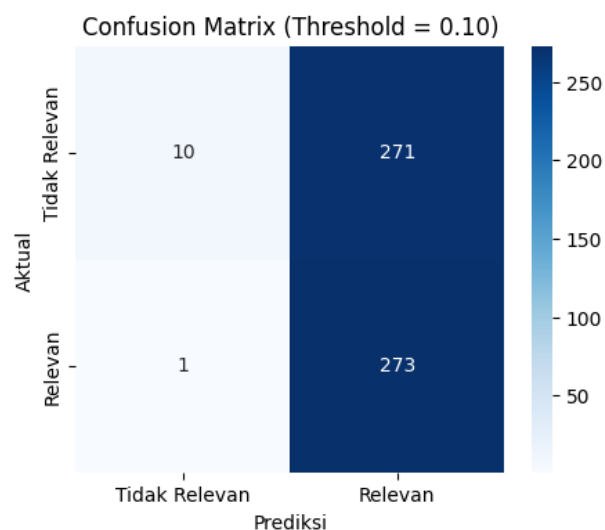
Proses evaluasi dilakukan menggunakan *Confusion Matrix* serta metrik evaluasi, yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. *Confusion Matrix* digunakan untuk mengetahui jumlah prediksi yang benar dan salah dalam setiap kategori, sedangkan *Accuracy*, *Precision*, *Recall*, dan *F1-Score* digunakan untuk mengukur kinerja sistem secara kuantitatif pada setiap skenario *threshold* yang telah ditentukan.

4.5.1 *Confusion Matrix*

Untuk mengevaluasi kinerja sistem dalam mengklasifikasikan tingkat relevansi antara judul dan transkrip video, dataset yang telah melalui proses pelabelan terlebih dahulu dibagi menjadi data training sebesar 80% atau 2.219 data dan data testing sebesar 20% atau 555 data. Data training digunakan sebagai dasar dalam proses pembentukan representasi semantik menggunakan Sentence BERT, sedangkan evaluasi kinerja sistem dilakukan pada data testing untuk mengukur kemampuan model dalam mengklasifikasikan data yang belum pernah digunakan pada tahap sebelumnya.

Evaluasi dilakukan menggunakan *Confusion Matrix* pada setiap nilai *threshold* yang telah ditentukan. *Confusion Matrix* digunakan untuk

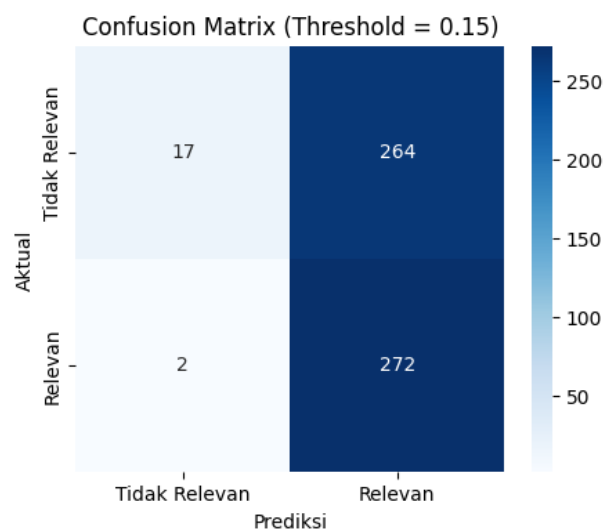
membandingkan hasil prediksi sistem dengan label aktual pada data testing sehingga dapat diketahui jumlah data yang diklasifikasikan dengan benar maupun yang mengalami kesalahan klasifikasi. Dari hasil perhitungan tersebut diperoleh nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) untuk masing-masing threshold. Gambar 4.20 hingga Gambar 4.24 menampilkan visualisasi Confusion Matrix pada data testing untuk setiap nilai threshold yang digunakan dalam penelitian ini. Visualisasi tersebut memberikan gambaran yang lebih jelas mengenai distribusi hasil klasifikasi serta pengaruh perubahan nilai threshold terhadap kemampuan sistem dalam membedakan data relevan dan tidak relevan.



Gambar 4. 20 *Confusion Matrix* pada *Threshold* 0.10

Berdasarkan Gambar 4.20, penggunaan threshold 0,10 menunjukkan bahwa sistem cenderung mengklasifikasikan sebagian besar data testing ke dalam kategori relevan. Hal ini terlihat dari nilai True Positive (TP) sebesar 273, yang menunjukkan bahwa hampir seluruh data relevan berhasil dikenali dengan benar

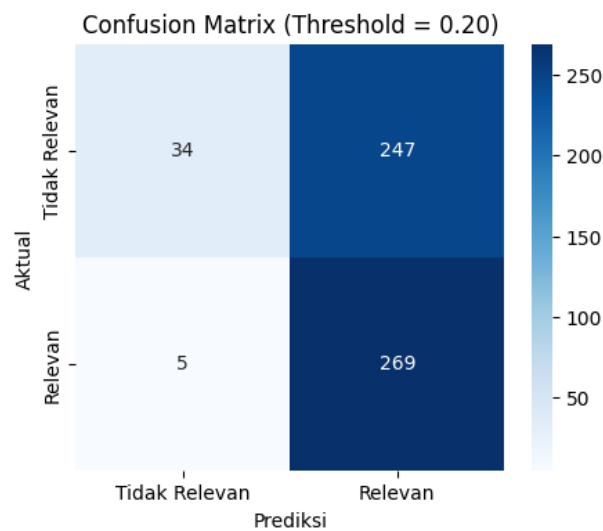
oleh sistem. Selain itu, nilai False Negative (FN) sebesar 1 menunjukkan bahwa hanya sedikit data relevan yang salah diklasifikasikan sebagai tidak relevan. Namun, nilai False Positive (FP) sebesar 271 mengindikasikan bahwa banyak data yang sebenarnya tidak relevan justru diprediksi sebagai relevan. Sementara itu, nilai True Negative (TN) sebesar 10 menunjukkan bahwa kemampuan sistem dalam mengenali data tidak relevan masih sangat rendah. Kondisi ini menunjukkan bahwa threshold 0,10 menghasilkan klasifikasi yang sangat longgar sehingga sistem memiliki kemampuan yang sangat baik dalam mendeteksi data relevan, tetapi kurang efektif dalam membedakan data yang tidak relevan. Yang mengakibatkan nilai recall menjadi sangat tinggi, namun diikuti oleh banyaknya kesalahan klasifikasi pada data tidak relevan.



Gambar 4. 21 *Confusion Matrix* pada *Threshold* 0.15

Pada threshold 0,15 sistem masih cenderung mengklasifikasikan data sebagai relevan. Hal ini terlihat dari nilai True Positive (TP) yang tinggi yaitu 272, yang menunjukkan bahwa sebagian besar data relevan berhasil dideteksi oleh

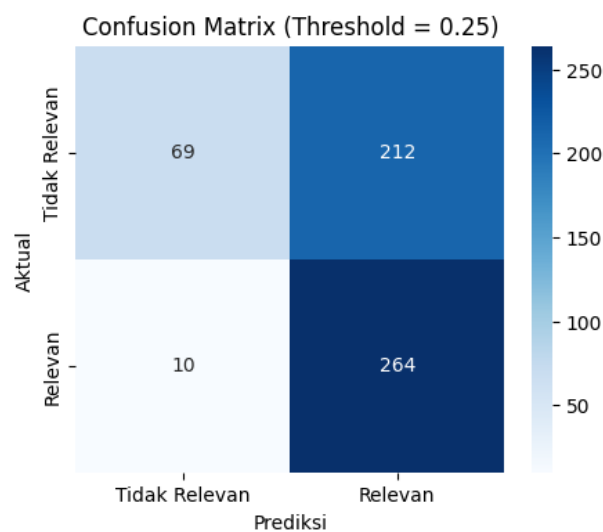
sistem. Nilai False Negative (FN) sebesar 2 menunjukkan bahwa hanya sangat sedikit data relevan yang terlewat. Namun nilai False Positive (FP) yang masih cukup tinggi yaitu 264 mengindikasikan bahwa banyak data tidak relevan diklasifikasikan sebagai relevan. Sementara nilai True Negative (TN) sebesar 17 menunjukkan bahwa kemampuan sistem dalam mengenali data tidak relevan masih terbatas. Kondisi ini menunjukkan bahwa threshold 0,15 masih menghasilkan klasifikasi yang cenderung bias ke arah relevan.



Gambar 4. 22 *Confusion Matrix* pada *Threshold* 0.20

Pada threshold 0,20 sistem masih mampu mendeteksi sebagian besar data relevan dengan baik. Hal ini terlihat dari nilai True Positive (TP) sebesar 269 yang menunjukkan bahwa sebagian besar data relevan berhasil diklasifikasikan dengan benar. Nilai False Negative (FN) sebesar 5 menunjukkan bahwa hanya sebagian kecil data relevan yang salah diklasifikasikan sebagai tidak relevan. Nilai False Positive (FP) sebesar 247 mengindikasikan bahwa masih terdapat cukup banyak data tidak relevan yang diklasifikasikan sebagai relevan. Nilai True Negative (TN)

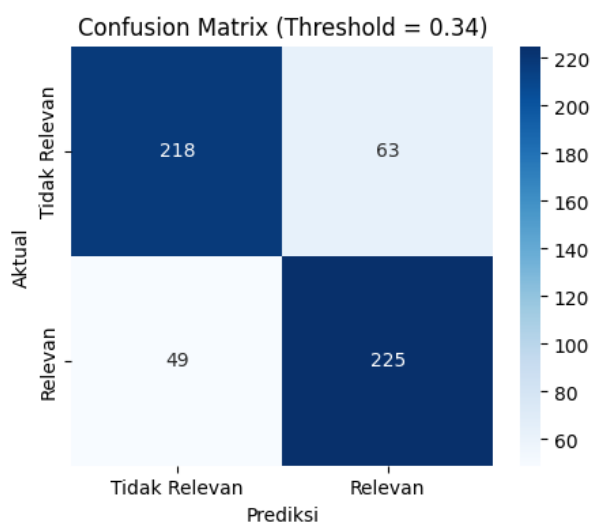
sebesar 34 menunjukkan bahwa sistem mulai mampu mengenali data tidak relevan dengan lebih baik. Kondisi ini menunjukkan bahwa threshold 0,20 menghasilkan proses klasifikasi yang lebih seimbang di mana sistem tetap dapat mendeteksi sebagian besar data relevan sekaligus mulai mengurangi kesalahan klasifikasi pada data tidak relevan.



Gambar 4. 23 *Confusion Matrix* pada *Threshold* 0.25

Pada threshold 0,25 sistem masih mampu mendeteksi sebagian besar data relevan dengan baik. Hal ini terlihat dari nilai True Positive (TP) sebesar 264 yang menunjukkan bahwa sebagian besar data relevan berhasil diklasifikasikan dengan benar. Nilai False Negative (FN) sebesar 10 menunjukkan bahwa terdapat sejumlah kecil data relevan yang tidak berhasil dikenali oleh sistem dan diklasifikasikan sebagai tidak relevan. Nilai False Positive (FP) sebesar 212 mengindikasikan bahwa masih terdapat data tidak relevan yang diklasifikasikan sebagai relevan. Nilai True Negative (TN) sebesar 69 menunjukkan bahwa sistem semakin mampu mengenali data tidak relevan dengan benar. Kondisi ini menunjukkan bahwa threshold 0,25

menghasilkan proses klasifikasi yang lebih ketat sehingga kemampuan sistem dalam membedakan data relevan dan tidak relevan menjadi lebih baik meskipun terdapat peningkatan jumlah data relevan yang tidak terdeteksi.



Gambar 4. 24 *Confusion Matrix* pada *Threshold* 0.34 (*Mean*)

Pada threshold 0,34 sistem menunjukkan kemampuan yang lebih seimbang dalam mengklasifikasikan data relevan dan tidak relevan. Hal ini terlihat dari nilai True Positive (TP) sebesar 225 yang menunjukkan bahwa sebagian besar data relevan masih berhasil dikenali dengan benar oleh sistem. Nilai False Negative (FN) sebesar 49 menunjukkan bahwa terdapat sejumlah data relevan yang diklasifikasikan sebagai tidak relevan. Nilai False Positive (FP) sebesar 63 mengindikasikan bahwa jumlah data tidak relevan yang salah diklasifikasikan sebagai relevan semakin berkurang. Nilai True Negative (TN) sebesar 218 menunjukkan bahwa sistem mampu mengenali data tidak relevan dengan sangat baik. Kondisi ini menunjukkan bahwa penggunaan threshold ini menghasilkan klasifikasi yang lebih seimbang antara kemampuan mendeteksi data relevan dan

kemampuan menghindari kesalahan klasifikasi pada data tidak relevan sehingga distribusi prediksi menjadi lebih proporsional dibandingkan skenario *threshold* yang lebih rendah.

Hasil perhitungan pada setiap *threshold* menunjukkan bahwa peningkatan nilai *threshold* menyebabkan sistem menjadi lebih selektif dalam melakukan klasifikasi. *Threshold* yang rendah cenderung menghasilkan kemampuan deteksi data relevan yang sangat tinggi, namun dengan tingkat kesalahan klasifikasi yang besar pada data tidak relevan. Sedangkan *threshold* yang lebih tinggi mampu mengurangi kesalahan tersebut, tetapi dengan konsekuensi meningkatnya jumlah data relevan yang tidak terdeteksi. Oleh karena itu pemilihan nilai *threshold* yang tepat menjadi faktor krusial dalam menentukan keseimbangan antara kemampuan sistem dalam mendeteksi data relevan atau *recall* dan kemampuan dalam menghindari kesalahan klasifikasi atau *precision*.

4.5.2 Hasil Evaluasi dan Perbandingan Kinerja

Berdasarkan hasil *confusion matrix* yang diperoleh pada setiap nilai *threshold*, selanjutnya dilakukan perhitungan metrik evaluasi untuk mengukur kinerja sistem secara kuantitatif. Metrik yang digunakan dalam penelitian ini meliputi *accuracy*, *precision*, *recall*, dan *F1-score*, yang dihitung berdasarkan nilai yang di dapatkan dari *confusion matrix*.

Hasil perhitungan metrik evaluasi untuk masing-masing nilai *threshold* ditunjukkan pada Tabel 4.12.

Tabel 4. 12 Hasil Perhitungan Metrik Evaluasi

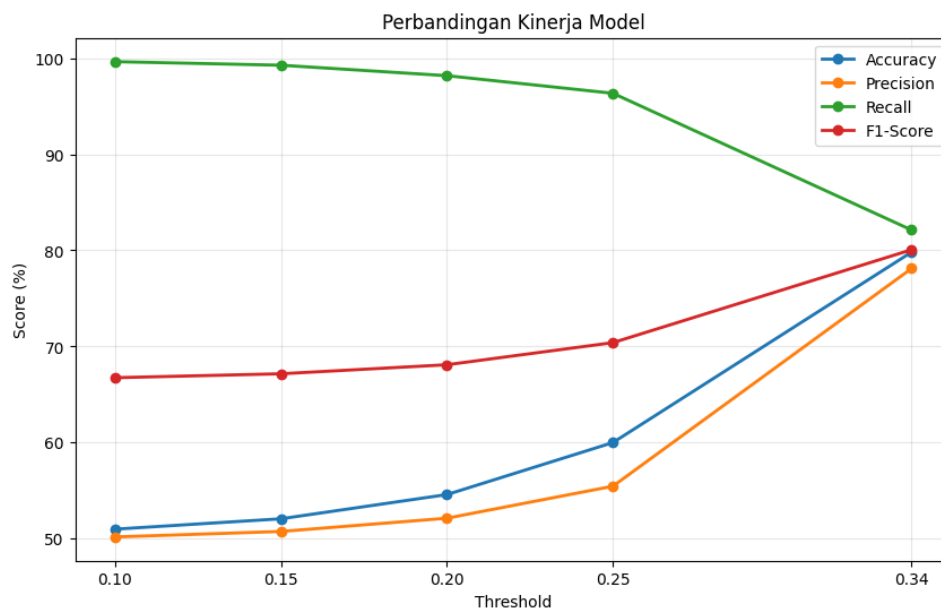
Threshold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.10	50.99	50.18	99.64	66.75
0.15	52.07	50.75	99.27	67.16
0.20	54.59	52.13	98.18	68.10
0.25	60.00	55.46	96.35	70.40
0,34 (Mean)	79.82	78.12	82.12	80.07

Berdasarkan Tabel 4.12 terlihat adanya hubungan yang saling memengaruhi antara nilai accuracy, precision, recall, dan F1-Score pada setiap skenario threshold. Recall yang tinggi menunjukkan kemampuan sistem dalam menemukan data relevan sedangkan precision menunjukkan ketepatan sistem dalam memberikan label relevan. Pada nilai threshold yang rendah sistem cenderung menghasilkan recall yang sangat tinggi karena hampir seluruh data relevan berhasil terdeteksi. Namun kondisi tersebut diikuti oleh precision yang relatif rendah akibat banyaknya data tidak relevan yang ikut diklasifikasikan sebagai relevan.

Seiring meningkatnya nilai threshold sistem menjadi lebih selektif dalam menentukan relevansi data. Dampaknya precision dan accuracy meningkat karena jumlah kesalahan klasifikasi berkurang tetapi recall mengalami penurunan karena sebagian data relevan tidak lagi terdeteksi. Kondisi ini menunjukkan adanya trade-off antara precision dan recall di mana peningkatan ketepatan klasifikasi umumnya diikuti oleh berkurangnya kemampuan sistem dalam menemukan seluruh data relevan.

F1-Score digunakan untuk mengukur keseimbangan antara precision dan recall. Semakin seimbang kedua metrik tersebut semakin tinggi nilai F1-Score yang dihasilkan. Pada penelitian ini threshold mean menghasilkan nilai accuracy,

precision, recall, dan F1-Score yang relatif seimbang dibandingkan skenario lainnya. Hal ini menunjukkan bahwa threshold mean mampu memberikan kompromi terbaik antara kemampuan sistem dalam mendeteksi data relevan dan ketepatan klasifikasi sehingga dipilih sebagai threshold terbaik pada penelitian ini.



Gambar 4. 25 Grafik Perbandingan Kinerja Model

Berdasarkan grafik pada Gambar 4.25 terlihat bahwa perubahan nilai threshold memberikan pengaruh yang signifikan terhadap kinerja sistem dalam mengklasifikasikan tingkat relevansi antara judul dan transkrip. Pada threshold yang rendah yaitu 0,10 hingga 0,25 nilai recall berada pada tingkat yang sangat tinggi sehingga sistem mampu mendeteksi hampir seluruh data relevan. Namun kondisi tersebut diikuti oleh nilai precision dan accuracy yang relatif rendah sehingga masih banyak data tidak relevan yang diklasifikasikan sebagai relevan.

Seiring meningkatnya nilai threshold nilai precision, accuracy, dan F1-Score cenderung mengalami peningkatan sedangkan nilai recall mengalami penurunan. Hal ini menunjukkan adanya trade-off antara precision dan recall di mana sistem menjadi lebih selektif dalam menentukan data relevan sehingga kesalahan klasifikasi pada data tidak relevan berkurang tetapi sebagian data relevan tidak lagi terdeteksi.

Nilai threshold mean sebesar 0,34 menghasilkan kinerja terbaik dengan nilai accuracy sebesar 79,82%, precision sebesar 78,12%, recall sebesar 82,12%, dan F1-Score sebesar 80,07%. Meskipun nilai recall lebih rendah dibandingkan threshold yang lebih rendah nilai tersebut masih berada pada tingkat yang baik dan diimbangi oleh peningkatan yang signifikan pada accuracy dan precision. Selain itu nilai F1-Score tertinggi menunjukkan bahwa threshold mean mampu memberikan keseimbangan terbaik antara kemampuan sistem dalam mendeteksi data relevan dan ketepatan klasifikasi. Oleh karena itu threshold mean dipilih sebagai threshold optimal pada penelitian ini karena menghasilkan performa klasifikasi yang paling baik secara keseluruhan.

4.6 Analisis Hasil

Berdasarkan hasil pengujian dan evaluasi yang telah dilakukan pada berbagai nilai threshold, sistem yang dikembangkan menggunakan metode *Sentence BERT* dan *Cosine Similarity* menunjukkan kemampuan yang baik dalam mengukur tingkat kemiripan antara judul dan transkrip secara semantik. Nilai evaluasi yang diperoleh menunjukkan bahwa sistem mampu mengidentifikasi

hubungan relevansi antara dua teks meskipun masih terdapat beberapa kesalahan klasifikasi pada beberapa pasangan data.

Hasil evaluasi menunjukkan bahwa penggunaan threshold mean sebesar 0,34 menghasilkan kinerja terbaik dengan nilai accuracy sebesar 79,82%, precision sebesar 78,12%, recall sebesar 82,12%, dan F1-Score sebesar 80,07%. Nilai F1-Score yang paling tinggi menunjukkan bahwa threshold mean mampu memberikan keseimbangan terbaik antara precision dan recall sehingga sistem dapat membedakan data relevan dan tidak relevan dengan lebih optimal.

Hasil analisis juga menunjukkan bahwa nilai *threshold* yang rendah cenderung menghasilkan nilai recall yang sangat tinggi namun diikuti oleh nilai precision dan accuracy yang rendah. Kondisi ini menunjukkan bahwa sistem cenderung terlalu mudah dalam menentukan relevansi sehingga banyak data tidak relevan diklasifikasikan sebagai relevan. Sebaliknya peningkatan nilai threshold membuat sistem menjadi lebih selektif yang ditunjukkan oleh meningkatnya nilai precision, accuracy, dan F1-Score. Dengan demikian pemilihan nilai threshold memiliki peran penting dalam menentukan keseimbangan antara kemampuan mendeteksi data relevan dan ketepatan klasifikasi yang dihasilkan sistem.

Tabel 4. 13 Analisis Perbandingan Kinerja Berdasarkan *Threshold*

<i>Threshold</i>	Kelebihan	Kekurangan
0.10	Recall sangat tinggi (99,64%)	Precision dan accuracy rendah
0.15	Recall sangat tinggi (99,27%)	Banyak false positive
0.20	Recall tinggi (98,18%)	Precision masih relatif rendah
0.25	F1-Score meningkat, accuracy lebih baik	Recall mulai menurun
0.3446 (Mean)	Accuracy, Precision, dan F1-Score tertinggi	Recall lebih rendah dibanding threshold lain

Berdasarkan Tabel 4.14, terlihat bahwa setiap nilai threshold menghasilkan karakteristik kinerja yang berbeda. *Threshold* yang rendah, yaitu 0,10 dan 0,15, menghasilkan nilai recall yang sangat tinggi karena sistem mampu mendeteksi hampir seluruh data relevan. Namun, kondisi tersebut diikuti oleh banyaknya data tidak relevan yang ikut diklasifikasikan sebagai relevan, sehingga nilai precision dan accuracy masih relatif rendah. Pada threshold 0,20 dan 0,25, sistem mulai menunjukkan peningkatan precision dan accuracy karena proses klasifikasi menjadi lebih selektif, meskipun nilai recall mulai mengalami penurunan.

Threshold mean sebesar 0,34 memberikan kinerja terbaik secara keseluruhan dengan menghasilkan nilai Accuracy, Precision, dan F1-Score tertinggi dibandingkan threshold lainnya. Hasil ini menunjukkan bahwa threshold mean mampu memberikan keseimbangan yang lebih baik antara ketepatan klasifikasi dan kemampuan mendeteksi data relevan. Meskipun nilai recall lebih rendah dibandingkan threshold yang lebih kecil, nilainya masih berada pada tingkat yang baik, sehingga threshold mean dipilih sebagai threshold yang paling optimal dalam penelitian ini.

Tabel 4. 14 Contoh Analisis Kesalahan Klasifikasi dengan *Threshold Mean*

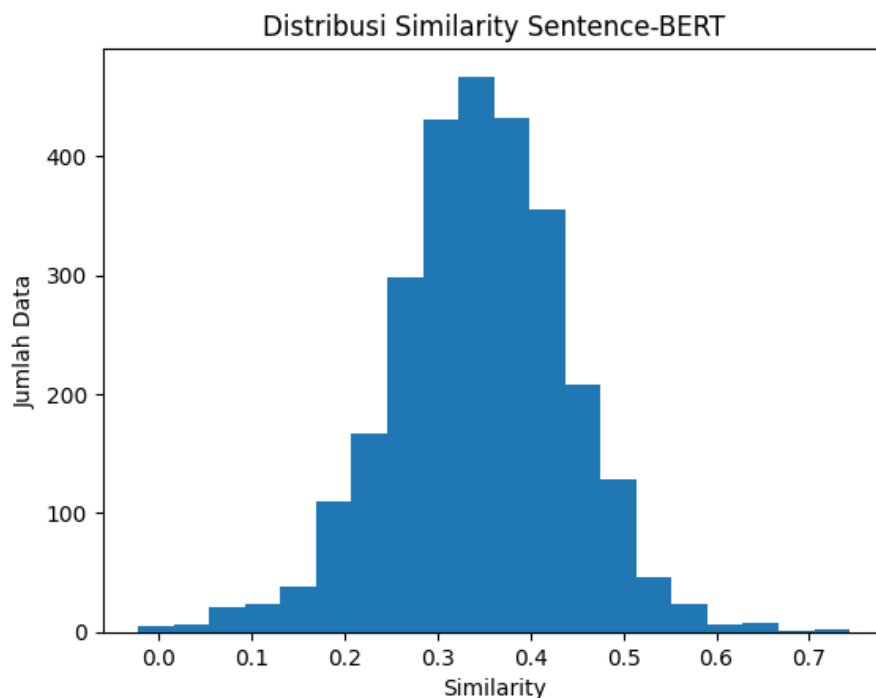
No	Judul	Transkrip	<i>Similarity</i>	Prediksi	Kondisi
1.	Boah Pngen Lulus Kuliah Biar Diumumin di Masjid SUCI PLAYGROUND #8	Bismillahirrahman irrahim. Asalamualaikum warahmatullahi	0,270527	0 (Tidak Relevan)	True Negative
2.	STAND UP RIO DUMATUBUN: GAK SEMUA BARANG COCOK DIJUAL DI	UBSI mana tepuk tangannya dong buat acara ini?	0,296468	0 (Tidak Relevan)	False Negative

	TIKTOK SUCI PLAYGROUND #8	Alhamdulillah ya bisa			
3.	VIRZA DISORAKIN BOHONG SAMA PENONTON SUCI PLAYGROUND #7	Asalamualaikum warahmatullah i wabarakatuh. Waalaikumsalam warahmatullah i wabarakatuh. Keren kalian	0,269726	0 (Tidak Relevan)	True Negative
4.	KEHIDUPAN RIZKY TEGUH DI JAKARTA TAK SEINDAH BAYANGANNYA SUCI PLAYGROUND #7	Asalamualaikum warahmatullah i wabarakatuh. Waalaikumsalam. Kenalin nama saya Rizki Teguh	0,429099	0 (Tidak Relevan)	False Negative
5.	CARA RAHMET MENGUKUR TINGKAT KEBAHAGIAAN MAHASISWA SUCI PLAYGROUND #7	Asalamualaikum warahmatullah i wabarakatuh. BSE. Waalaikumsalam warahmatullah i wabarakatuh. Luar biasa	0,403429	0 (Tidak Relevan)	False Negative
.....					
2774 .	Agungprov: Nyamar Jadi Sapu Masjid - SUCI 8 (OPEN MIC)	saya Cepu ya asalamualaikum warahmatullah i wabarakatuh selamat malam perkenalkan	0,253936	0 (Tidak Relevan)	True Negative

Tabel 4.15 menampilkan contoh analisis hasil klasifikasi menggunakan threshold mean sebesar 0,34. Tabel tersebut memuat judul, sebagian transkrip, nilai similarity hasil perhitungan Sentence BERT dan Cosine Similarity, prediksi sistem,

serta kondisi klasifikasi yang dihasilkan. Data dengan nilai *similarity* di bawah *threshold* dikategorikan sebagai tidak relevan (label 0). Pada tabel terlihat beberapa data termasuk kategori *True Negative*, yaitu data yang memang tidak relevan dan berhasil diklasifikasikan sebagai tidak relevan oleh sistem. Kondisi ini menunjukkan bahwa penggunaan *threshold mean* mampu meningkatkan kemampuan sistem dalam mengenali data yang tidak memiliki hubungan semantik yang kuat antara judul dan transkrip.

Terdapat beberapa data yang termasuk kategori *False Negative*, yaitu data yang sebenarnya relevan tetapi diklasifikasikan sebagai tidak relevan karena nilai *similarity* berada di bawah *threshold mean* yang ditetapkan. Kondisi ini menunjukkan bahwa penggunaan *threshold mean* membuat proses klasifikasi menjadi lebih ketat, sehingga mampu mengurangi kesalahan pada data tidak relevan, namun menyebabkan sebagian data relevan tidak berhasil terdeteksi. Hal ini tercermin dari meningkatnya nilai *accuracy* dan *precision*, tetapi diikuti dengan penurunan nilai *recall*.



Gambar 4. 26 Distribusi *Similarity Sentence BERT*

Berdasarkan Gambar 4.26, distribusi nilai *similarity Sentence BERT* menunjukkan pola yang menyerupai distribusi normal dengan konsentrasi terbesar berada pada rentang 0,25 hingga 0,45. Puncak distribusi terlihat di sekitar nilai 0,34, yang menunjukkan bahwa sebagian besar pasangan judul dan transkrip memiliki tingkat kemiripan semantik pada kategori sedang. Kondisi ini menunjukkan bahwa mayoritas data berada di sekitar nilai rata-rata *similarity*, sehingga pemilihan *threshold* memiliki pengaruh yang besar terhadap hasil klasifikasi. Nilai *threshold* yang terlalu rendah cenderung mengklasifikasikan sebagian besar data sebagai relevan, sedangkan *threshold* yang terlalu tinggi dapat menyebabkan banyak data relevan tidak terdeteksi. Hasil distribusi ini mendukung penggunaan *threshold* mean sebesar 0,34 sebagai batas klasifikasi, karena nilai tersebut berada pada pusat

penyebaran data dan menghasilkan kinerja terbaik berdasarkan hasil evaluasi. Hal ini dibuktikan dengan perolehan nilai accuracy sebesar 79,06%, precision sebesar 78,02%, recall sebesar 80,13%, dan F1-score sebesar 79,06%, yang merupakan nilai tertinggi dibandingkan skenario threshold lainnya. Dengan demikian, threshold mean dapat dianggap sebagai nilai yang paling optimal untuk menentukan tingkat relevansi antara judul dan transkrip pada penelitian ini.

4.7 Integrasi Islam

Penelitian mengenai pengukuran tingkat relevansi antara judul dan transkrip video memiliki keterkaitan dengan nilai-nilai dalam ajaran islam, khususnya yang berkaitan dengan kejujuran, ketepatan informasi, dan tanggung jawab dalam penyampaian informasi kepada orang lain.

Salah satu prinsip penting dalam Islam adalah melakukan verifikasi terhadap informasi sebelum menerimanya. Hal ini dijelaskan dalam QS. Al-Hujurat ayat 6 yang menekankan pentingnya memastikan kebenaran suatu informasi agar tidak menimbulkan kesalahan dalam pengambilan keputusan.

يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْحَبُوا عَلَىٰ مَا فَعَلْتُمْ

نَدِمِينَ ﴿٦﴾ (الحجرات/49: 6)

“Wahai orang-orang yang beriman, jika datang kepadamu seorang fasik membawa suatu berita, maka periksalah dengan teliti, agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu.” (Q.S. Al-Hujurat: 6)

Tafsir ayat ini menjelaskan bahwa setiap informasi yang diterima tidak boleh langsung dipercaya tanpa melalui proses pemeriksaan yang cermat, terutama

apabila sumber informasi tersebut belum dapat dipastikan kebenarannya. Kehatian dalam menerima berita menjadi hal yang sangat penting agar tidak menimbulkan kesalahan dalam pengambilan keputusan yang dapat merugikan pihak lain. Sikap tergesa-gesa dalam menerima dan menyebarkan informasi berpotensi menimbulkan penyesalan di kemudian hari, terutama jika informasi tersebut ternyata tidak benar. Ayat ini memberikan pedoman kepada umat Islam untuk selalu bersikap teliti, kritis, dan bertanggung jawab dalam menyikapi setiap informasi yang diterima, sehingga dapat menghindari dampak negatif akibat kesalahan dalam memahami atau menyampaikan suatu berita (Kemenag 2023).

Keselarasan antara judul dan isi suatu konten berkaitan dengan prinsip kejujuran dalam menyampaikan informasi, di mana setiap informasi yang disampaikan harus didasarkan pada kebenaran dan tidak menyesatkan pihak lain. Ketidaksesuaian antara judul dan isi dapat menimbulkan persepsi yang salah, sehingga bertentangan dengan nilai kejujuran dalam Islam. Allah SWT berfirman dalam Surah Al-Ahzab ayat 70,

يَا أَيُّهَا الَّذِينَ آمَنُوا اتَّقُوا اللَّهَ وَقُولُوا قَوْلًا سَدِيدًا ﴿٧٠﴾ (الاحزاب/33: 70)

“Wahai orang-orang yang beriman, bertakwalah kamu kepada Allah dan ucapkanlah perkataan yang benar.” (Q.S. Al-Ahzab: 70)

Tafsir ayat ini menjelaskan bahwa setiap orang beriman diperintahkan untuk menjaga ketakwaan kepada Allah dengan cara menyampaikan perkataan yang benar, jujur, dan tepat sasaran. Perkataan yang disampaikan tidak hanya harus sesuai dengan fakta, tetapi juga harus disampaikan secara jelas dan tidak menimbulkan kesalahpahaman. Hal ini menunjukkan bahwa kebenaran dalam

komunikasi bukan hanya terletak pada isi, tetapi juga pada cara penyampaiannya agar tidak menyesatkan pihak lain. Oleh karena itu ayat ini menekankan pentingnya tanggung jawab dalam berbicara, sehingga setiap informasi yang disampaikan dapat memberikan manfaat dan tidak menimbulkan dampak negatif bagi orang lain (Kemenag 2023).

Aspek tanggung jawab dalam menerima dan menyampaikan informasi juga menjadi bagian penting dalam ajaran Islam, di mana setiap individu dituntut untuk tidak mengikuti atau menyebarkan informasi tanpa dasar pengetahuan yang jelas serta harus bersikap hati-hati dalam menilai kebenaran suatu informasi. Allah SWT berfirman dalam Surah Al-Isra ayat 36,

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهِ عِلْمٌ إِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ أُولَٰئِكَ كَانَ عَنْهُ مَسْئُولًا ﴿٣٦﴾ (الاسراء/17):
(36)

“Janganlah engkau mengikuti sesuatu yang tidak kau ketahui. Sesungguhnya pendengaran, penglihatan, dan hati nurani, semua itu akan diminta pertanggung jawabannya.” (Q.S. Al-Isra: 36)

Tafsir ayat ini menjelaskan bahwa setiap individu dilarang untuk mengatakan atau mengikuti sesuatu tanpa didasari oleh pengetahuan yang jelas dan dapat dipertanggung jawabkan. Seseorang tidak diperbolehkan mengklaim mengetahui, melihat, atau mendengar sesuatu yang sebenarnya tidak dialami secara langsung, karena hal tersebut dapat menimbulkan kesalahan dan menyesatkan orang lain. Pendengaran, penglihatan, dan hati dipandang sebagai amanah yang diberikan oleh Allah SWT, sehingga penggunaannya harus dijaga dengan penuh tanggung jawab. Setiap informasi yang diterima maupun disampaikan akan dimintai pertanggungjawaban, sehingga diperlukan sikap kehati-hatian dan

kejujuran dalam menyikapi serta menyampaikan suatu informasi agar tidak menimbulkan dampak negatif (Kemenag 2023).

Keterkaitan antara nilai-nilai tersebut dengan penelitian ini terlihat pada upaya sistem dalam mengukur kesesuaian antara judul dan isi transkrip video. Proses tersebut dapat dipahami sebagai bentuk verifikasi informasi secara otomatis yang membantu dalam mengidentifikasi apakah suatu konten disampaikan secara sesuai atau tidak. Penelitian ini merupakan salah satu upaya untuk menerapkan nilai-nilai yang terkandung dalam ayat tersebut, khususnya terkait pentingnya melakukan pemeriksaan dan verifikasi informasi sebelum informasi tersebut diterima atau disebarluaskan.

Nilai-nilai kejujuran, verifikasi informasi, dan tanggung jawab yang terkandung dalam ajaran Islam memberikan landasan etis dalam pengembangan dan pemanfaatan teknologi, khususnya dalam pengolahan informasi. Pendekatan yang digunakan dalam penelitian ini diharapkan dapat mendukung penyebaran informasi yang lebih akurat dan dapat dipercaya di era digital.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, tingkat kecocokan antara judul dan transkrip video dapat diukur menggunakan representasi semantik berbasis *Sentence BERT* dan perhitungan *Cosine Similarity*. Proses pengukuran dilakukan melalui tahap preprocessing, tokenisasi, pembentukan embedding, mean pooling, dan perhitungan nilai similarity. Nilai kemiripan yang diperoleh digunakan untuk menentukan tingkat relevansi antara judul dan transkrip sehingga sistem mampu mengukur kecocokan berdasarkan kesamaan makna secara semantik.

Hasil penelitian menunjukkan bahwa sistem yang dikembangkan mampu mengklasifikasikan relevansi judul dan transkrip video ke dalam kategori relevan dan tidak relevan secara objektif. Dari 2774 data yang digunakan, diperoleh 1405 data tidak relevan dan 1369 data relevan. Berdasarkan hasil evaluasi, threshold mean sebesar 0,34 menghasilkan performa terbaik dengan nilai accuracy 79,82%, precision 78,12%, recall 82,12%, dan F1-score 80,07%. Hasil tersebut menunjukkan bahwa kombinasi *Sentence BERT* dan *Cosine Similarity* efektif digunakan untuk mengukur tingkat kecocokan antara judul dan transkrip video.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan beragam agar kemampuan model dalam mengukur relevansi dapat diuji pada kondisi yang lebih

kompleks dan representatif. Proses pelabelan data juga dapat melibatkan validator manusia untuk menghasilkan *ground truth* yang lebih akurat. Pengembangan penelitian dapat dilakukan dengan membandingkan performa *Sentence BERT* terhadap model representasi semantik lainnya, seperti IndoBERT atau model transformer terbaru, serta menerapkan metode optimasi untuk menentukan nilai *threshold* yang lebih optimal. Hasil penelitian ini juga berpotensi untuk diterapkan pada sistem nyata, seperti pencarian video, rekomendasi konten, atau validasi kesesuaian antara judul dan isi video secara otomatis.

DAFTAR PUSTAKA

- Adelpina, Dea, Salma Salsabila Prissa, and Rinaldi Supriadi. 2024. "Teknik Penerjemahan Google Translate Dan Subtitle YouTube Indonesia Arab; Serial Nussa 'Rarra Sakit.'" *Aphorisme: Journal of Arabic Language, Literature, and Education* 5(1):101–16. doi:10.37680/aphorisme.v5i1.5207.
- Albab, M. Ulil, Yohana Karuniawati P., and Mohammad Nur Fawaiq. 2023. "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic." *Jurnal Transformatika* 20(2):1–12. doi:10.26623/transformatika.v20i2.5374.
- Alwasi'i, Tubagus. 2025. "Generator Makna Lirik Lagu Menggunakan Transformer (Studi Kasus : Lirik Lagu Drake)." *Jurnal Transformatika* 22(1):1–12. doi:10.26623/transformatika.v22i1.5744.
- Ayuningtyas, Nimas, and Wiyli Yustanti. 2024. "Semi-Supervised Learning Pada Pelabelan Dalam Klasifikasi Multi-Label Data Teks." *Journal of Informatics and Computer Science (JINACS)* 6(01):240–48. doi:10.26740/jinacs.v6n01.p240-248.
- Dawud, Moh. 2025. "Klasifikasi Sentimen Terhadap Program Makan Bergizi Gratis Menggunakan Metode Logistic Regression." *Frontiers in Veterinary Science* 13(1):1–16.
- Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1(Mlm):4171–86.
- Fathuddin, Muhammad Abdul Hafizh, Eka Prakarsa Mandyartha, and Afina Lina Nurlaili. 2025. "Penerapan *Sentence BERT* Dan Cosine Similarity Untuk Pencarian Semantik Dokumen Skripsi Dalam Format PDF." *Ranah Research : Journal of Multidisciplinary Research and Development* 8(1):322–37. doi:10.38035/rj.v8i1.1865.
- Febianto Agiharta, Kafka, Bernard Renaldy Suteja, and Ayub Mewati. 2024. "Penerapan *Sentence BERT* Untuk Similaritas Kompetensi Pekerjaan Dan Mata Kuliah." *Jurnal Teknik Informatika Dan Sistem Informasi* 10(3):1–19.
- Ferdiansyah, Ari. 2025. "Sistem Penilaian Esai Otomatis Menggunakan Algoritma *Sentence Bidirectional Encoder Representation From Transformers (Sbert)* Pada Sekolah Menengah Kejuruan Mekanik Cibinong Bogor." 1–9.
- Habibi, Roni, and Naufal Dekha Widana. 2025. "Penerapan Algoritma K-Nearest Neighbors Untuk Deteksi Serangan Network Flood Berbasis Supervised Learning." *Jurnal Sains, Nalar, Dan Aplikasi Teknologi Informasi* 4(2):129–37. doi:10.20885/snati.v4.i2.38599.

- Hakiki, Ilhan, Firas Atqiya, and Ahmad Suryan. 2025. "Implementasi Model *Sentence BERT* Untuk Deteksi Plagiarisme Pada Karya Tulis Ilmiah Psikologi Berbahasa Indonesia." *Cerdika: Jurnal Ilmiah Indonesia* 5(12):3109–25. doi:10.59141/cerdika.v5i12.2889.
- Hartatik, Hartatik, and Andri Syafrianto. 2025. "Penerapan Model *Sentence BERT* Untuk Sistem Rekomendasi Buku Berbasis Konten Di Perpustakaan Digital." *Jurnal Dialektika Informatika (Detika)* 6(1):12–19. doi:10.24176/detika.v6i1.15916.
- Holis, Rahul Marcellino, Pradita Eko Prasetyo Utomo, and Benedika Ferdian Hutabarat. 2025. "Semantic FAQ Chatbot Using SBERT (*Sentence BERT*) and Cosine Similarity for Academic Services." *Brilliance: Research of Artificial Intelligence* 5(2):915–22. doi:10.47709/brilliance.v5i2.7027.
- Kadzim, Moh. Musa Al. 2025. "Analisis Sentimen Berbasis Aspek Terhadap Program Makan Bergizi Gratis Di Media Sosial X Menggunakan Metode Support Vector Machine."
- Kemenag. 2023. *Qur'an Kementerian Agama Republik Indonesia, Tafsir Al-Qur'an Dan Terjemahan*. Jakarta: Lajnah Pentashihan Mushaf Al-Qur'an.
- Kho, Andy, and Felliks F. Tampinongkol. 2025. "Analisis Sentimen Pengguna Aplikasi STEAM Dengan Algoritma Naive Bayes." *Ranah Research : Journal of Multidisciplinary Research and Development* 7(6):4620–27. doi:10.38035/rrj.v7i6.1803.
- Lahitani, Alfirna, Ulfi Saidata Aesyi, Noviana Wulandari, and Bagas Dwi Santosa. 2022. "Cosine Similarity Untuk Mengukur Tingkat Kesadaran Pada Topik Software Security Berbasis Teks Komentar Di Media Sosial Youtube." *Jurnal Sains Dan Informatika* 8(2):107–15. doi:10.34128/jsi.v8i2.535.
- Larasakti, Dickna Niken. 2023. "Analisis Sentimen Komentar Video Youtube Dengan Metode K-Nearest Neighbor." 9(5):132–42.
- Maulidya Prastita Syah, Ajeng Puspa Wardani, Mohammad Idhom, and Trimono. 2025. "Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks." *Data Sciences Indonesia (DSI)* 5(1):47–59. doi:10.47709/dsi.v5i1.6021.
- Nisfi, Azza Muhammad, Albi Salman Bachtiar, Rudianto Rudianto, and Cucu Ika Agustyaningrum. 2025. "Analisis Perbandingan Algoritma Klasifikasi Machine Learning Dalam Pendeteksian Risiko Stunting Pada Balita Di Desa Sukahati." *JlIP - Jurnal Ilmiah Ilmu Pendidikan* 8(9):10562–68. doi:10.54371/jiip.v8i9.9077.
- Nurwanda, Nurwanda, Nana Suarna, and Willy Prihartono. 2024. "Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore." *JATI (Jurnal Mahasiswa Teknik Informatika)* 8(2):1841–46. doi:10.36040/jati.v8i2.8469.

- Ompusunggu, Martin, A. Rafi, Paringgom Iwari, Silvia Azahrani, Devina Anin Safitri, Annisa Cahyani Surya, Nadinda Aurora, Ardika Satria, Christyan Tamaro Nadeak, and Ade Lailani. 2025. "Implementasi Sentence Transformer Berbasis Bert Untuk Edukasi Dan Informasi Tentang Hiv/Aids." *Prosiding Seminar Nasional Sains Dan Teknologi Seri III Fakultas Sains Dan Teknologi* 2(1):607–14.
- Pitriani, Dian Sa'adillah Maylawati, and Yana Aditia Gerhana. 2025. "Deteksi Generatif Teks Pada Penilaian Otomatis Tes Esai Berbahasa Indonesia Menggunakan IndoBERT." *Jurnal Edukasi Dan Penelitian Informatika* 11(2).
- Rohimim, Nurma Fadlilah. 2025. "Peringkasan Teks Ekstraktif Pada Artikel Berita Bahasa Indonesia Menggunakan Metode Bert Dan Cosine Similarity." 43–44.
- Supriyono Supriyono; Aji Prasetya Wibawa; Suyono Suyono; Fachrul Kurniawan. 2025. *Indonesian Stand-Up Comedy Transcription Dataset*. Malang: Mendeley Data.
- Susilowati, Rifha. 2021. "Clickbait Pada Headline Youtube (Analisis Isi Pesan Terhadap Judul-Judul Konten Clickbait Di Channel Bobon Santoso Edisi Konten Tebarunya)." 6.
- Utami, Alike Putri. 2024. "Proses Morfologis Pada Judul Berita Di Youtube Tvonenews." 3(9):2312–27.
- Wahyuni, Widya, Tri Putri Lestari, Milla Apriliana, and Riyang Gumelta. 2025. "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments." *Journal of Applied Informatics and Computing* 9(4):1964–71. doi:10.30871/jaic.v9i4.9754.
- Widiantoro, Albertus, Dwiyooga Mustafid Sanjaya, and Ridwan. 2024. *Pengantar NLP Dan Topik Model*.