EVALUASI KESESUAIAN TEKS PADA WEBSITE BUDAYA-INDONESIA.ORG MENGGUNAKAN METODE TF-IDF DENGAN VARIASI *RULE-BASED VALIDATION*

SKRIPSI

Oleh :
NAZLA SABILA ROSYAD
NIM. 220605110145

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

EVALUASI KESESUAIAN TEKS PADA WEBSITE BUDAYA-INDONESIA.ORG MENGGUNAKAN METODE TF-IDF DENGAN VARIASI *RULE-BASED VALIDATION*

SKRIPSI

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

Oleh :
NAZLA SABILA ROSYAD
NIM. 220605110145

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

HALAMAN PERSETUJUAN

EVALUASI KESESUAIAN TEKS PADA WEBSITE BUDAYA-INDONESIA.ORG MENGGUNAKAN METODE TF-IDF DENGAN VARIASI *RULE-BASED VALIDATION*

SKRIPSI

Oleh :
NAZLA SABILA ROSYAD
NIM. 220605110145

Telah Diperiksa dan Disetujui untuk Diuji:

Tanggal: 12 Juni 2026

Pembimbing I,	Pembimbing II,
Roro Inda Melani, MT, M.Sc	Ahmad Fahmi Karami, M.Kom
NIP. 19780925 200501 2 008	NIP. 19870909 202012 1 001

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

EVALUASI KESESUAIAN TEKS PADA WEBSITE BUDAYA-INDONESIA.ORG MENGGUNAKAN METODE TF-IDF DENGAN VARIASI *RULE-BASED VALIDATION*

SKRIPSI

Oleh :
NAZLA SABILA ROSYAD
NIM. 220605110145

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	: Dr. Ir. Fresy Nugroho, ST., MT, IPM NIP. 19710722 201101 1 001	( )
Anggota Penguji I	: Ashri Shabrina Afrah, M.T NIP. 19900430 202012 2 003	( )
Anggota Penguji II	: Roro Inda Melani, M.T, M.Sc NIP. 19780925 200501 2 008	( )
Anggota Penguji III	: Ahmad Fahmi Karami, M.Kom NIP. 19870909 202012 1 001	( )

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	: Nazla Sabila Rosyad
NIM	: 220605110145
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Evaluasi Kesesuaian Text pada Website Budaya-Indonesia.org Menggunakan Metode TF-IDF dengan Variasi *Rule-Based Validation*

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, 13 Juni 2026
Yang membuat pernyataan,

Nazla Sabila Rosyad
NIM. 220605110145

MOTTO

"Kapan seseorang benar-benar mati? Ketika mereka tertembak peluru? Tidak. Ketika mereka menderita penyakit? Tidak. Ketika mereka memakan sup jamur beracun? Tidak! Seseorang benar-benar mati hanya ketika mereka dilupakan!"

(Dr.Hiriluk, One Piece)

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur, karya ini penulis persembahkan kepada:

Alm. Ayah, Ibu, dan kakak-kakak tercinta

Atas doa, dukungan, dan kasih sayang tanpa henti yang selalu menjadi sumber kekuatan dan motivasi.

Seluruh keluarga besar

Yang senantiasa memberikan dukungan dan doa dalam setiap langkah penulis.

Dosen pembimbing dan para pengajar

Atas bimbingan, arahan, serta ilmu yang sangat berarti bagi penulis.

Teman-teman Infinity Teknik Informatika 2022

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

Dan akhirnya, untuk diri sendiri

Yang terus berusaha, melangkah, dan bertahan hingga mencapai titik ini.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahirabbil'alamin, segala puji dan syukur senantiasa penulis haturkan kepada Allah SWT. yang telah melimpahkan rahmat serta bimbingan-Nya, sehingga penulis berhasil menuntaskan skripsi berjudul "EVALUASI KESESUAIAN TEKS PADA WEBSITE BUDAYA-INDONESIA.ORG MENGGUNAKAN METODE TF-IDF DENGAN VARIASI *RULE-BASED VALIDATION*" dengan baik. Shalawat dan salam penulis curahkan kepada Rasulullah Muhammad SAW., panutan agung umat manusia yang telah menuntun kita dari masa kebodohan menuju cahaya kebenaran, agama Islam, serta ilmu pengetahuan yang hingga kini masih kita rasakan manfaatnya.

Skripsi ini disusun sebagai salah satu persyaratan untuk meraih gelar Sarjana Komputer di Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim Malang. Selama proses pengerjaannya, penulis senantiasa mendapat bantuan, motivasi, dan dukungan doa dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin mengucapkan 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. Roro Inda Melani, M.T, M.Sc, selaku Dosen Pembimbing I, yang dengan kesabaran dan dedikasinya membimbing penulis selama penyusunan skripsi.
5. Ahmad Fahmi Karami, M.Kom, selaku Dosen Pembimbing II, yang senantiasa memberikan arahan dan masukan yang sangat berharga dalam pengerjaan skripsi ini.
6. Dr. Ir. Fresy Nugroho, ST., MT, IPM., dan Ashri Shabrina Afrah, M.T., 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. Alm. Makhsun Hasan, Bchk, S.Pd dan Ibu Mukhofifah, orang tua tercinta, serta Rahmad Agung Febriansyah, S.Si dan Muhammad Iqbal Fahmi, M.Si, kakak saya, yang selalu menjadi pilar kekuatan dengan doa, cinta, dan dukungan tanpa henti.
9. Rekan seperjuangan “2RU”, yang beranggotakan Rheza, Maul, Izzan, Fiki, Lana. Karya kecil ini kupersembahkan untuk kalian yang selalu menjadi sumber semangatku. Terima kasih sudah mau berbagi tawa dan air mata selama ini.

10. Teman solid “PDI” yang beranggotakan Ridiey dan Syafri. Yang membersamai baik ketika berada dikala senang maupun sedih.
11. Seluruh keluarga, teman, sahabat, dan kerabat penulis yang tidak dapat penulis sebutkan satu per satu, yang turut memberikan bantuan, semangat, dukungan, dan doa untuk penulis.

Dan terakhir tak lupa terimakasih kepada diri sendiri, yang dengan tekad kuat mampu melewati setiap tantangan, rintangan, dan kesulitan selama perjalanan ini, membuktikan bahwa kerja keras dan usaha tidak pernah sia-sia. Penulis menyadari bahwa penelitian dalam tugas skripsi ini masih jauh dari kata sempurna dan memiliki berbagai keterbatasan. Oleh karena itu, dengan penuh kerendahan hati, penulis membuka diri untuk menerima kritik dan saran yang membangun dari para pembaca guna menjadi bahan evaluasi dan pengembangan di masa mendatang. Penelitian ini juga memiliki potensi untuk dilanjutkan dan dikembangkan lebih jauh dalam penelitian berikutnya, sehingga dapat melengkapi kekurangan yang ada. Penulis berharap, karya ini tidak hanya memberikan manfaat bagi pembaca, tetapi juga dapat berkontribusi positif bagi masyarakat secara luas.

Wassalamu’alaikum Warahmatullahi Wabarakatuh.

Malang, 22 Juni 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Rosyad, Nazla Sabila. 2026. **Evaluasi Kesesuaian Teks pada Website Budaya-Indonesia.org Menggunakan Metode TF-IDF dengan Variasi Rule-Based Validation.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Roro Inda Melani, M.T, M.Sc; (II) Ahmad Fahmi Karami, M.Kom.

.

Kata kunci: TF-IDF, Rule-Based Validation, Cosine Similarity, Kesesuaian Teks, Text Mining, Budaya Digital.

Pelestarian budaya digital melalui platform *crowdsourcing* seperti budaya-indonesia.org menimbulkan tantangan pada kualitas data, khususnya ketidaksesuaian semantik antara judul dan deskripsi entri budaya. Penelitian ini mengevaluasi tingkat kesesuaian teks menggunakan metode TF-IDF (*Term Frequency–Inverse Document Frequency*) yang dikombinasikan dengan *rule-based validation* pada 58.604 entri data dari situs budaya-indonesia.org. Tahapan penelitian mengikuti kerangka CRISP-DM yang meliputi *data preparation* dengan proses *case folding, tokenizing, stopword removal,* dan *stemming*, dilanjutkan dengan perhitungan *cosine similarity* berbasis bobot TF-IDF untuk mengklasifikasikan kesesuaian ke dalam tiga kategori: Sangat Sesuai, Cukup Sesuai, dan Tidak Sesuai. Hasil implementasi menunjukkan bahwa pendekatan hybrid TF-IDF dan *rule-based* mampu meningkatkan sensitivitas sistem dalam mendeteksi keterkaitan kontekstual yang tidak tertangkap oleh TF-IDF saja, ditandai dengan peningkatan kategori Sangat Sesuai dari 6% menjadi 13,9% dan peningkatan kategori Cukup Sesuai dari 53,3% menjadi 79,7%, sehingga secara keseluruhan sistem mampu meningkatkan proporsi data yang terklasifikasi sesuai secara signifikan. Evaluasi kualitatif pada 15 skenario percobaan yang dirancang dalam tiga level data uji dan lima tema budaya menunjukkan seluruh skenario berhasil mengalami peningkatan kategori dengan tingkat keberhasilan 100%.

ABSTRACT

Rosyad, Nazla Sabila. 2026. **Evaluation of Text Relevance on the Budaya-Indonesia.org Website Using the TF-IDF Method with Variations in Rule-Based Validation.** Thesis. Computer Science Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University, Malang. Advisors: (I) Roro Inda Melani, M.T., M.Sc.; (II) Ahmad Fahmi Karami, M.Kom.

Keywords: TF-IDF, Rule-Based Validation, Cosine Similarity, Text Relevance, Text Mining, Digital Culture.

.

The preservation of digital culture through crowdsourcing platforms such as budaya-indonesia.org poses challenges regarding data quality, particularly semantic mismatches between the titles and descriptions of cultural entries. This study evaluates the level of text alignment using the TF-IDF (Term Frequency–Inverse Document Frequency) method combined with rule-based validation on 58,604 data entries from the budaya-indonesia.org website. The research phases followed the CRISP-DM framework, which included data preparation through case folding, tokenization, stopword removal, and stemming, followed by the calculation of TF-IDF-weighted cosine similarity to classify alignment into three categories: Highly Aligned, Moderately Aligned, and Not Aligned. The implementation results show that the hybrid TF-IDF and rule-based approach was able to improve the system's sensitivity in detecting contextual relationships not captured by TF-IDF alone, as indicated by an increase in the "Highly Relevant" category from 6% to 13.9% and an increase in the "Moderately Relevant" category from 53.3% to 79.7%, thereby enabling the system to significantly increase the proportion of correctly classified data overall. A qualitative evaluation of 15 experimental scenarios designed across three levels of test data and five cultural themes showed that all scenarios successfully improved their classification categories with a 100% success rate.

ملخص

روزاد، نازلا سابيلا. 2026. **تقييم ملاءمة النصوص على موقع Budaya-Indonesia.org باستخدام طريقة TF-IDF مع تنويعات في التحقق القائم على القواعد**. أطروحة تخرج. برنامج هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية، مالانج. المشرفان: (I) رورو إندا ميلاني، ماجستير في الهندسة، ماجستير في العلوم؛ (II) أحمد فهمي كارامي، ماجستير في علوم الحاسوب.

الكلمات المفتاحية: TF-IDF، التحقق القائم على القواعد، تشابه جيب التمام، ملاءمة النصوص، استخراج النصوص، الثقافة الرقمية.

يُشكل الحفاظ على الثقافة الرقمية من خلال منصات *التعهيد الجماعي* مثل budaya-indonesia.org تحديات تتعلق بجودة البيانات، ولا سيما عدم التوافق الدلالي بين عناوين ووصف المدخلات الثقافية. تقيّم هذه الدراسة مستوى توافق النصوص باستخدام طريقة TF-IDF (*تكرار المصطلح – تكرار الوثيقة العكسي*) مقترنةً *بالتحقق القائم على القواعد* على 58,604 مدخلاً من موقع budaya-indonesia.org. تتبع مراحل البحث إطار عمل CRISP-DM الذي يشمل *إعداد البيانات* من خلال عمليات *«توحيد الحالات» (case folding)*، و*«التقطيع إلى رموز» (tokenizing)*، و*«إزالة الكلمات الممنوعة» (stopword removal)*، و*«الاستخلاص» (stemming)*، تليها حساب *«تشابه جيب التمام» (cosine similarity)* استنادًا إلى أوزان TF-IDF لتصنيف التوافق إلى ثلاث فئات: «متوافق جدًا»، و«متوافق إلى حد ما»، و«غير متوافق». أظهرت نتائج التنفيذ أن النهج الهجين بين TF-IDF *والقواعد القاعدية* قادر على تحسين حساسية النظام في الكشف عن الارتباطات السياقية التي لا يلتقطها TF-IDF وحده، وهو ما يتجلى في ارتفاع نسبة فئة «مطابقة جدًا» من 6% إلى 13,9% وارتفاع نسبة فئة «مطابقة إلى حد ما» من 53,3% إلى 79,7%، وبالتالي، تمكن النظام بشكل عام من زيادة نسبة البيانات المصنفة بشكل صحيح بشكل ملحوظ. أظهر التقييم النوعي لـ 15 سيناريو تجريبي تم تصميمها على ثلاثة مستويات من بيانات الاختبار وخمسة مواضيع ثقافية أن جميع السيناريوهات نجحت في تحسين التصنيف بنسبة نجاح بلغت 100%.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pelestarian budaya merupakan salah satu aspek penting dalam menjaga identitas nasional Indonesia yang memiliki keragaman budaya sangat tinggi (Putro et al., 2022). Di era digital, upaya pelestarian tersebut tidak lagi hanya dilakukan secara fisik, tetapi juga melalui proses digitalisasi data budaya agar dapat diakses secara luas dan berkelanjutan. Transformasi digital ini memungkinkan masyarakat dan peneliti untuk berkontribusi dalam memperkaya basis pengetahuan budaya melalui platform daring. Menurut Gagliardi & Artese (2024), digitalisasi warisan budaya berbasis data terbuka dan multilingual menjadi pilar penting dalam memperkuat pemahaman lintas daerah dan bahasa. Di Indonesia, salah satu upaya besar dalam mendigitalisasi kekayaan budaya dilakukan melalui situs budaya-indonesia.org, yang memuat lebih dari 58.000 entri budaya yang bersumber dari kontribusi masyarakat secara daring (Wibowo & Salim, 2024).

Situs budaya-indonesia.org menjadi wujud nyata *crowdsourcing* pengetahuan budaya, di mana setiap individu dapat menambahkan data, deskripsi, dan narasi terkait unsur budaya lokal (Wibowo & Salim, 2024). Keberadaan situs ini memiliki nilai strategis bagi pelestarian budaya nasional karena menjadi wadah kolaborasi pengetahuan antara masyarakat dan institusi akademik. Namun, di sisi lain, mekanisme kontribusi terbuka ini menimbulkan tantangan pada aspek kualitas data, terutama pada indikasi kesesuaian semantik antara judul dan deskripsi.

Banyak entri yang menampilkan perbedaan makna, ketidaktepatan konteks, atau pengulangan informasi, yang dapat mengurangi kredibilitas dan akurasi basis data budaya nasional. Permasalahan ini relevan dengan pandangan Wibowo & Salim (2024) yang menekankan bahwa partisipasi publik dalam pengelolaan data budaya perlu diimbangi dengan mekanisme verifikasi untuk menjaga keandalan informasi. Oleh karena itu, situs budaya-indonesia.org menjadi objek penelitian yang penting karena mencerminkan pergeseran paradigma pelestarian budaya ke arah partisipatif digital, namun sekaligus menunjukkan kebutuhan akan evaluasi otomatis terhadap kesesuaian dan kualitas teks yang diunggah publik.

Penggunaan pendekatan *crowdsourcing* dalam pengumpulan data juga menimbulkan tantangan serius terkait validitas dan konsistensi informasi. Menurut Kaldeli et al. (2021) serta Ba et al. (2025), data kolaboratif sering menghadapi masalah kualitas seperti kesalahan penulisan, serta ketidaksesuaian antara judul dan deskripsi. Pada web budaya-indonesia.org pun ditemukan kasus di mana judul "Tari Piring" justru disertai deskripsi mengenai "alat musik tradisional." Contoh ini menunjukkan lemahnya verifikasi otomatis dalam sistem *crowdsourced* data budaya. Wibowo & Salim (2024) menegaskan bahwa tingkat indikasi konsistensi semantik sangat menentukan kredibilitas data budaya digital karena ketidaksesuaian dapat menghambat proses pelestarian dan pengambilan keputusan.

Dampak dari data yang tidak sesuai tidak hanya bersifat teknis, tetapi juga langsung memengaruhi kualitas representasi pengetahuan budaya pada platform digital. Ketika entri budaya memuat informasi yang tidak sesuai antara judul dan deskripsi, akurasi konten menjadi terganggu dan menurunkan keandalan data secara

keseluruhan (Kaldeli et al., 2021; Wibowo & Salim, 2024). Ketidaktepatan ini terutama menyulitkan proses verifikasi otomatis serta evaluasi kesesuaian teks pada sistem seperti budaya-indonesia.org. Zheng et al. (2024) dalam studinya mengenai pemrosesan bahasa alami pada warisan budaya takbenda menjelaskan bahwa kualitas data teks memiliki pengaruh langsung terhadap efektivitas model analisis budaya digital. Dengan demikian, perlu dilakukan evaluasi kesesuaian antara elemen-elemen teks penting seperti judul dan deskripsi, agar setiap entri budaya dapat merepresentasikan konsep dan konteksnya secara konsisten. Pendekatan ini juga sejalan dengan penelitian Julians et al. (2024) yang menegaskan pentingnya penerapan *text mining* untuk menemukan pola dan anomali indikasi kesesuaian semantik dalam konten budaya kolaboratif.

TF-IDF *(Term Frequency–Inverse Document Frequency)* merupakan pendekatan statistik yang mengukur relevansi suatu kata dalam konteks dokumen, dan telah terbukti efektif dalam berbagai penelitian pengukuran kemiripan teks (Amrulloh & Adam, 2021). Selain itu, penelitian ini menerapkan variasi *rule-based system* untuk melengkapi keterbatasan metode TF-IDF dalam menangkap makna linguistik yang tidak selalu tercermin secara numerik. Variasi ini berupa aturan-aturan eksplisit yang mendeteksi indikator kesesuaian, seperti keterkaitan kata judul, konteks budaya, dan panjang deskripsi, yang digunakan untuk menyesuaikan kategori hasil tanpa mengubah nilai similarity dasar. Selanjutnya, hasil model TF-IDF tanpa *rule-based* dibandingkan dengan model yang menggunakan *rule-based* untuk menilai pengaruh penambahan aturan terhadap ketepatan klasifikasi kesesuaian teks.

Sistem berbasis aturan *(rule-based system)* adalah pendekatan dalam *Natural Language Processing* (NLP) yang menggunakan kumpulan aturan linguistik yang dirancang secara eksplisit untuk melakukan analisis, pengelompokan, maupun validasi terhadap data teks (Pulungan, 2024). Dengan demikian, penelitian ini tidak hanya berkontribusi terhadap peningkatan kualitas data budaya digital, tetapi juga memberikan landasan bagi pengembangan sistem verifikasi otomatis yang dapat diterapkan dalam pelestarian budaya nasional di masa depan.

Selain berlandaskan pendekatan ilmiah, penelitian ini juga sejalan dengan nilai-nilai Islam yang menekankan pentingnya menjaga dan mengenali kebudayaan sebagai bagian dari tanda-tanda kebesaran Allah. Dalam Al-Qur'an, Allah berfirman:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ۝٦

"Wahai orang-orang yang beriman, apabila datang kepadamu seorang fasik membawa suatu berita, maka periksalah dengan teliti (tabayyun), agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya, yang menyebabkan kamu menyesal atas perbuatanmu itu." (QS. Al-Hujurat [49]: 6)

Berdasarkan tafsir Tafsir Al-Misbah, ayat ini mengajarkan pentingnya melakukan verifikasi terhadap informasi yang diterima sebelum mengambil keputusan atau menyebarkannya kepada orang lain. Perintah *tabayyun* bertujuan agar seseorang tidak terburu-buru dalam mempercayai suatu berita yang belum jelas kebenarannya, karena hal tersebut dapat menimbulkan kesalahan penilaian, kerugian bagi pihak lain, dan penyesalan di kemudian hari. Nilai ini sejalan dengan tujuan penelitian yang berupaya memverifikasi kesesuaian antara judul dan

deskripsi data budaya menggunakan metode TF-IDF. Dengan menerapkan prinsip *tabayyun* dalam konteks ilmiah, penelitian ini menekankan pentingnya ketelitian dan keakuratan data sebagai dasar menjaga kebenaran informasi budaya digital Indonesia. Penelitian ini menjadi penting karena hasilnya dapat berkontribusi langsung dalam meningkatkan kualitas dan kredibilitas data budaya nasional, sehingga warisan budaya yang terdigitalisasi tetap autentik, terpercaya, dan bernilai edukatif bagi masyarakat.

1.2 Rumusan Masalah

Rumusan masalah ini berfokus pada evaluasi kesesuaian antara judul dan deskripsi data budaya pada situs budaya-indonesia.org menggunakan metode TF-IDF dan *rule-based* validation. Permasalahan dalam penelitian ini adalah bagaimana mengukur dan mengevaluasi tingkat kesesuaian antara judul dan deskripsi data budaya menggunakan metode TF-IDF, serta bagaimana pengaruh penerapan *rule-based* validation terhadap hasil klasifikasi kesesuaian teks dalam menilai kualitas informasi budaya digital di Indonesia.

1.3 Batasan Masalah

1. Data penelitian dari situs budaya-indonesia.org 58.709 sejak November 2025 diambil dari situs budaya-indonesia.org sebagai sumber utama informasi budaya digital.
2. Analisis difokuskan pada dua kolom data teks, yaitu judul dan deskripsi, tanpa melibatkan kolom lain.

3. Penelitian ini memfokuskan pada pemrosesan data berbasis teks. Entri yang hanya berisi tautan tanpa konten tekstual tidak termasuk dalam analisis.

1.4 Tujuan Masalah

Tujuan dari penelitian ini adalah untuk mengukur tingkat kesesuaian antara kolom judul dan deskripsi pada data budaya yang terdapat di situs budaya-indonesia.org menggunakan metode TF-IDF *(Term Frequency–Inverse Document Frequency)*. Melalui penerapan metode ini, penelitian berupaya mengidentifikasi seberapa besar tingkat kesesuaian antar teks sehingga dapat diketahui konsistensi informasi dalam setiap entri budaya. Selain itu, penelitian ini mencoba menerapkan variasi *rule-based validation* sebagai mekanisme evaluatif untuk menyesuaikan hasil pengukuran kemiripan berdasarkan indikator linguistik dan konteks budaya, sehingga proses penilaian kesesuaian menjadi lebih objektif dan kontekstual. Hasil penelitian diharapkan dapat menjadi dasar dalam evaluasi dan peningkatan kualitas data budaya digital, khususnya pada aspek relevansi dan validitas deskripsi terhadap judul, serta memberikan kontribusi bagi pengelola data budaya nasional dalam pengembangan sistem verifikasi yang lebih efisien, transparan, dan mendukung pelestarian budaya Indonesia melalui basis data yang akurat dan terstruktur.

1.5 Manfaat Penelitian

1. **Manfaat Akademik**

   Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang *text mining* dan *natural*

language processing (NLP), khususnya dalam penerapan NLP pada konteks kebudayaan Indonesia di website budaya-indonesia.org. Melalui analisis kesesuaian antara judul dan deskripsi, penelitian ini memperluas pemahaman mengenai bagaimana metode statistik seperti TF-IDF dapat digunakan untuk menilai relevansi data berbasis teks dalam ranah budaya. Hasil penelitian ini juga dapat dijadikan referensi bagi penelitian selanjutnya yang berfokus pada peningkatan kualitas data budaya atau penerapan teknik NLP pada data non-formal dan *crowdsourced*.

2. **Manfaat Praktis**

   Secara praktis, penelitian ini bermanfaat sebagai dasar verifikasi bagi pengelola data budaya nasional, seperti pengembang situs budaya-indonesia.org atau lembaga kebudayaan yang mengelola konten digital. Dengan adanya hasil analisis kesesuaian teks, sistem pengelolaan data dapat memiliki indikator awal untuk menilai apakah suatu entri budaya valid atau perlu diperiksa ulang secara manual. Hal ini akan membantu meningkatkan efisiensi proses kurasi data serta memperkuat akurasi informasi budaya yang disajikan kepada publik.

3. **Manfaat Teknis**

   Dari sisi teknis, penelitian ini menawarkan pendekatan implementatif yang sederhana namun efektif, yaitu penerapan metode TF-IDF yang dikombinasikan dengan *rule-based* system untuk mendeteksi kesesuaian teks antara dua kolom utama. Pendekatan ini dapat dijadikan acuan dalam pengembangan sistem evaluasi otomatis yang mampu menangani data

dalam jumlah besar tanpa memerlukan model pembelajaran yang kompleks. Selain itu, hasil penelitian ini juga memberikan gambaran tentang potensi pemanfaatan metode serupa dalam sistem validasi data digital di berbagai domain lainnya.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Penelitian mengenai penerapan *text mining* dan metode TF-IDF (*Term Frequency–Inverse Document Frequency*) telah banyak dilakukan dalam berbagai konteks, terutama untuk mengukur kesamaan teks dan meningkatkan kualitas data digital. Berikut adalah beberapa penelitian yang memiliki relevansi tinggi dengan topik ini.

Amrulloh & Adam (2021) melakukan penelitian berjudul *Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF-IDF* yang bertujuan untuk menghitung tingkat kemiripan antarjudul tugas akhir menggunakan pendekatan *cosine similarity*. Hasil penelitian menunjukkan bahwa metode TF-IDF efektif dalam mendeteksi kesamaan konteks antarjudul meskipun memiliki perbedaan kata. Studi ini menjadi dasar penting dalam penelitian ini karena memiliki kemiripan pada proses pengukuran *semantic similarity* antar teks pendek, seperti antara judul dan deskripsi data budaya.

Selanjutnya, penelitian oleh Julians et al. (2024) dengan judul *Analisis Konten Budaya Kolaboratif Berbasis Grounded Theory Menggunakan Text Mining* menunjukkan penerapan *text mining* dalam konteks budaya Indonesia. Penelitian ini berfokus pada analisis konten budaya hasil kontribusi masyarakat *(crowdsourced content)* untuk menggali makna budaya yang tersembunyi di dalam teks. Pendekatan ini memperlihatkan bahwa teknik pengolahan teks tidak hanya

relevan dalam ranah teknis, tetapi juga mampu mendukung pelestarian budaya secara digital.

Penelitian lain oleh Putro et al. (2022) dalam karya berjudul *Eksplorasi Metode Crowdsourcing dalam Upaya Pengarsipan Musik melalui Perancangan Web-based Director* membahas penerapan metode *crowdsourcing* untuk pengarsipan budaya musik. Studi ini menyoroti pentingnya validasi dan verifikasi data hasil kontribusi publik agar arsip digital tetap berkualitas. Hal ini sejalan dengan kebutuhan penelitian ini, yakni memastikan kualitas dan kesesuaian data teks pada situs budaya-indonesia.org.

Terakhir, penelitian Wibowo & Salim (2024) berjudul *Pengaruh Crowdsourcing dan Representasi Koleksi terhadap Kredibilitas Informasi pada Museum di Indonesia* membahas faktor-faktor yang memengaruhi keandalan dan kredibilitas data budaya. Mereka menekankan pentingnya representasi data yang akurat agar informasi budaya yang disajikan tetap sesuai dan dipercaya publik. Relevansi penelitian ini terletak pada upaya menjaga kualitas informasi budaya digital melalui evaluasi teks secara sistematis.

Berdasarkan hasil tinjauan keempat penelitian tersebut, dapat disimpulkan bahwa meskipun telah banyak penelitian yang mengangkat tema *text mining* dan pelestarian budaya digital, sebagian besar masih berfokus pada analisis konten, klasifikasi, atau pengarsipan data budaya. Penelitian ini menghadirkan celah penelitian *(research gap)* baru dengan fokus pada evaluasi kesesuaian antar kolom teks (judul dan deskripsi) menggunakan pendekatan TF-IDF serta didukung oleh

aturan berbasis linguistik (*rule-based system*) sederhana. Berikut merupakan penelitian terkait yang ada pada Tabel 2.1.

Tabel 2. 1 Penelitian Terkait

No	Peneliti & Tahun	Judul Penelitian	Metode	Hasil	Relevansi
1	Amrulloh & Adam (2021)	Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF-IDF	TF-IDF dan *Cosine similarity*	Menghasilkan sistem pencarian kemiripan teks yang efektif antarjudul tugas akhir.	Menjadi dasar pengukuran kesesuaian teks antara judul dan deskripsi pada data budaya.
2	Bashir et al. (2021)	*Subjective Answers Evaluation Using Machine Learning and Natural Language Processing*	NLP dan *Machine Learning* dengan kombinasi Word2Vec, *Word Mover's Distance* (WMD), *Cosine similarity*, TF-IDF, dan *Multinomial Naive Bayes* (MNB).	*Cosine similarity* menghasilkan akurasi sekitar 87% tanpa model dan 86% dengan model.	Membahas pengukuran kesamaan teks menggunakan TF-IDF dan *Cosine similarity* dalam evaluasi jawaban deskriptif.
3	Sari et al. (2021)	Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma	Text Mining, TF-IDF, dan *Cosine similarity*	Dari 50 abstrak yang diuji, 34 abstrak berhasil diklasifikasikan sesuai kategori sehingga memperoleh tingkat keberhasilan sebesar 73%.	Relevan karena sama-sama menggunakan TF-IDF dan *Cosine similarity* dalam mengukur kemiripan teks.
4	Julians, Manongga & Hendry (2024)	Analisis Konten Budaya Kolaboratif Berbasis Grounded Theory Menggunakan Text Mining	*Text Mining*, Grounded Theory	Mengungkap makna budaya dari data teks hasil kontribusi publik.	Relevan dalam konteks penerapan *text mining* untuk pelestarian budaya digital.
5	Putro, Larasati & Ratri (2022)	Eksplorasi Metode Crowdsourcing dalam Upaya Pengarsipan Musik melalui Perancangan Web-*based* Director	*Crowdsourcing*, Web-*based* System	Menunjukkan pentingnya validasi data budaya hasil kontribusi masyarakat.	Mendukung konsep verifikasi otomatis kualitas data budaya daring.

6	Wibowo & Salim (2024)	Pengaruh Crowdsourcing dan Representasi Koleksi terhadap Kredibilitas Informasi pada Museum di Indonesia	Studi Kualitatif, Analisis Representasi Data	Menemukan hubungan antara representasi data dan kredibilitas informasi budaya.	Menguatkan urgensi evaluasi kesesuaian teks untuk menjaga validitas data budaya nasional.
7	Chamira (2022)	Implementasi Metode Text Mining Frequency-Inverse Document Frequency (TF-IDF) untuk Monitoring Diskusi Online	TF-IDF, Text Mining	Menunjukkan TF-IDF mampu mendeteksi kesesuaian topik dan relevansi antar kalimat dalam forum daring.	Mendukung metode pengukuran kesesuaian teks antar kolom (judul dan deskripsi) menggunakan pendekatan serupa.
8	Widianto, Pebriyanto & Fitriyanti (2024)	Document Similarity Using Term Frequency–Inverse Document Frequency Representation and *Cosine similarity*	TF-IDF dan *Cosine similarity*	TF-IDF terbukti efektif dalam mengukur kesamaan dokumen dengan tingkat akurasi tinggi (93,6%).	Menjadi dasar ilmiah dalam pemilihan metode TF-IDF untuk mengevaluasi indikasi kesesuaian semantik pada data budaya digital.

Tabel 2.1 di atas menampilkan rangkuman beberapa penelitian terdahulu yang relevan dengan topik evaluasi kesesuaian teks dan pelestarian budaya digital. Enam penelitian tersebut mencakup dua arah utama:

1. Arah teknis, yaitu penerapan metode *Text Mining* dan TF-IDF untuk pengukuran kesamaan teks (Amrulloh & Adam, 2021; Chamira, 2022; Widianto et al., 2024).
2. Arah kontekstual, yaitu penerapan teknologi informasi dalam pengelolaan dan validasi data budaya digital (Julians et al., 2024; Putro et al., 2022; Wibowo & Salim, 2024).

Melalui tinjauan tersebut, dapat disimpulkan bahwa penelitian terdahulu telah banyak menggunakan TF-IDF untuk mendeteksi kesamaan teks, namun belum ada yang secara spesifik mengevaluasi indikasi kesesuaian semantik antara judul dan deskripsi pada data budaya daring. Oleh karena itu, penelitian ini mengisi celah tersebut dengan mengintegrasikan TF-IDF dan pendekatan *rule-based* sederhana untuk analisis kesesuaian judul dan isi pada situs budaya-indonesia.org.

2.2 Landasan Teori

Landasan teori merupakan dasar konseptual yang menjelaskan prinsip, konsep, serta penelitian terdahulu yang mendukung penelitian ini. Dalam konteks penelitian ini, objek kajian difokuskan pada situs budaya-indonesia.org atau Perpustakaan Digital Budaya Indonesia (PDBI), sebuah platform digital yang dibangun oleh *Indonesian Archipelago Culture Initiatives (IACI)* untuk menghimpun, mendokumentasikan, dan mendigitalisasi artefak budaya Nusantara. Situs ini memuat lima belas kategori artefak, seperti tarian, ritual, pakaian tradisional, permainan rakyat, hingga naskah kuno, yang dikumpulkan melalui sistem *crowdsourcing* memungkinkan masyarakat menjadi kontributor aktif dalam menambahkan dan memperbarui data budaya. Meskipun berperan penting sebagai repositori nasional dan sumber pengetahuan budaya yang terbuka, sistem partisipatif ini menimbulkan tantangan terhadap kualitas dan kesesuaian informasi yang diunggah publik. Oleh karena itu, penelitian ini didasari teori-teori mengenai pengolahan teks (*text mining*), konsep kesesuaian semantik, serta metode TF-IDF *(Term Frequency–Inverse Document Frequency)* sebagai pendekatan statistik

untuk menilai relevansi dan hubungan makna antar teks, guna mengevaluasi tingkat keakuratan dan validitas data budaya pada situs budaya-indonesia.org.

2.2.1 *Text Mining*

Text mining merupakan proses mengekstraksi informasi dan pengetahuan yang bermanfaat dari kumpulan data teks tidak terstruktur. Menurut Chamira (2022), *text mining* melibatkan serangkaian tahapan seperti pembersihan data *(text preprocessing)*, tokenisasi, penghapusan kata tidak penting *(filtering/stopword removal),* serta stemming atau lemmatisasi untuk mendapatkan representasi kata yang lebih bermakna. Proses ini memungkinkan sistem komputer mengenali pola linguistik dan indikasi semantik yang tersembunyi di dalam teks secara efisien.

Dalam konteks yang lebih luas, *text mining* berperan penting dalam bidang *Natural Language Processing* (NLP), karena berfungsi sebagai tahapan awal dalam memahami struktur dan makna bahasa alami. Widianto et al. (2024) menekankan bahwa integrasi *text mining* dan NLP memungkinkan analisis hubungan indikasi semantik antarteks melalui pengukuran kemiripan dan klasifikasi berbasis konteks. Selain itu, Chamira (2022) menunjukkan bahwa *text mining* mampu digunakan untuk mendeteksi kesesuaian topik, relevansi antar kalimat, serta konteks semantik pada dokumen digital.

Dalam penelitian ini, *text mining* berfungsi sebagai fondasi utama untuk menganalisis hubungan antara judul dan deskripsi pada data budaya Indonesia. Melalui penerapan metode TF-IDF *(Term Frequency–Inverse Document Frequency),* penelitian ini berusaha mengukur bobot relevansi kata secara statistik

untuk mengevaluasi sejauh mana indikasi kesesuaian semantik antar teks dalam dataset budaya daring. Pendekatan ini menjadi langkah awal untuk memastikan validitas representasi informasi budaya secara digital dan mendukung peningkatan kualitas data pada platform budaya-indonesia.org.

2.2.1.1 *Text Processing*

Text preprocessing merupakan tahapan awal dalam text mining yang bertujuan untuk membersihkan, menormalkan, dan mempersiapkan data teks agar dapat diolah oleh sistem komputer secara lebih efektif. Data teks yang diperoleh dari sumber digital umumnya masih mengandung berbagai bentuk noise, seperti perbedaan huruf kapital, tanda baca, angka, kata-kata umum yang tidak memiliki makna penting, serta variasi bentuk kata yang dapat memengaruhi hasil analisis. Oleh karena itu, preprocessing menjadi langkah penting untuk meningkatkan kualitas representasi teks dan akurasi proses analisis selanjutnya (Chamira, 2022).

Menurut Widianto et al., (2024), *text preprocessing* membantu mengurangi kompleksitas data dan meningkatkan konsistensi representasi linguistik sehingga hubungan semantik antarteks dapat dianalisis dengan lebih baik. Tahapan preprocessing yang umum digunakan dalam penelitian text mining meliputi *case folding, tokenizing, stopword removal, stemming, dan cleaning noise* (Widianto et al., 2024).

1. *Case Folding*

   Case folding merupakan proses mengubah seluruh huruf dalam teks menjadi huruf kecil *(lowercase).* Tahapan ini dilakukan untuk menghindari

perbedaan representasi kata yang sebenarnya memiliki makna sama, seperti "Tari", "tari", dan "TARI".

2. *Tokenizing*

   Tokenizing adalah proses memecah teks menjadi unit-unit kata atau token. Hasil tokenisasi digunakan sebagai dasar dalam proses analisis teks, termasuk perhitungan frekuensi kata dan pembobotan istilah.

3. *Stopword Removal*

   Stopword removal merupakan proses menghapus kata-kata umum yang sering muncul tetapi memiliki kontribusi makna yang rendah terhadap isi dokumen, seperti "dan", "yang", "di", atau "dari". Penghapusan stopword bertujuan meningkatkan fokus analisis pada kata-kata yang lebih informatif.

4. *Stemming*

   Stemming adalah proses mengubah kata berimbuhan menjadi bentuk dasarnya *(root word).* Sebagai contoh, kata "menari", "penari", dan "tarian" dapat direduksi menjadi kata dasar "tari". Proses ini membantu menyatukan variasi bentuk kata yang memiliki makna serupa.

5. *Cleaning Noise*

   Cleaning noise merupakan proses pembersihan karakter yang tidak relevan terhadap analisis teks, seperti tanda baca, simbol khusus, angka, atau karakter lain yang tidak memiliki nilai linguistik. Tahapan ini bertujuan menghasilkan data teks yang lebih bersih dan terstruktur.

Melalui serangkaian tahapan tersebut, teks dapat direpresentasikan secara lebih konsisten sehingga siap digunakan pada proses pembobotan kata dan

pengukuran kemiripan dokumen. Dalam penelitian ini, hasil *preprocessing* digunakan sebagai masukan pada metode TF-IDF untuk menghitung bobot kata dan menganalisis tingkat kesesuaian antara judul dan deskripsi data budaya Indonesia.

2.2.1.2 TF-IDF *(Term Frequency–Inverse Document Frequency)*

Metode TF-IDF *(Term Frequency–Inverse Document Frequency)* merupakan salah satu teknik representasi teks paling populer dalam bidang *text mining* dan *information retrieval*. Menurut Chamira (2022), TF-IDF berfungsi untuk menilai tingkat kepentingan suatu kata dalam sebuah dokumen dibandingkan dengan seluruh kumpulan dokumen *(corpus)*. Konsep ini terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) yang mengukur frekuensi kemunculan kata dalam dokumen, dan *Inverse Document Frequency* (IDF) yang memberikan penalti terhadap kata-kata yang terlalu sering muncul di banyak dokumen karena dianggap kurang informatif.

Secara matematis, nilai TF-IDF diperoleh melalui perkalian dua komponen berikut:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k \quad f_{k,d}} \; dan \; IDF(t) = \log \frac{N}{df_t} \qquad (2.\ 1)$$

Keterangan:
$f_{t,d}$ = jumlah kemunculan kata t pada dokumen d.
N = jumlah total dokumen dalam korpus.
df_t = jumlah dokumen yang mengandung kata t.

Nilai gabungan dari kedua komponen tersebut menghasilkan bobot akhir $w_{t,d} = TF(t,d) \times IDF(t)$ (Widianto et al., 2024).

Kelebihan TF-IDF terletak pada kesederhanaan dan efektivitasnya dalam mengidentifikasi kata penting dari teks tanpa memerlukan pelatihan data atau sumber daya besar. Menurut Widianto et al. (2024), TF-IDF tetap menjadi metode yang relevan karena mampu memberikan hasil akurat untuk pengukuran kesamaan dokumen dengan tingkat efisiensi tinggi, terutama saat dipadukan dengan *cosine similarity*. Sementara itu, Chamira (2022) membuktikan bahwa TF-IDF dapat digunakan untuk menganalisis kesesuaian topik dan relevansi antar kalimat pada data teks daring, memperkuat perannya dalam konteks evaluasi kesamaan semantik.

Meskipun demikian, TF-IDF memiliki keterbatasan dalam menangkap makna semantik yang mendalam, karena memperlakukan setiap kata secara independen tanpa mempertimbangkan konteks linguistik. Oleh sebab itu, beberapa penelitian seperti Julians et al. (2024) menyarankan agar analisis berbasis TF-IDF dapat dikombinasikan dengan pendekatan berbasis aturan (*rule-based*) atau konteks linguistik untuk meningkatkan interpretasi semantik. Dalam penelitian ini, TF-IDF dipilih karena kemampuannya memberikan bobot numerik yang objektif terhadap relevansi kata serta efisien untuk dataset besar seperti budaya-indonesia.org.

Hasil pembobotan TF-IDF menghasilkan representasi numerik dalam bentuk vektor untuk setiap teks. Namun, representasi tersebut belum secara langsung menunjukkan tingkat hubungan antar teks. Oleh karena itu, diperlukan metode pengukuran kemiripan untuk membandingkan vektor yang dihasilkan (Widianto et al., 2024). Dalam penelitian ini, *cosine similarity* digunakan sebagai metode pengujian untuk menghitung tingkat kesesuaian antara vektor judul dan deskripsi yang telah direpresentasikan menggunakan TF-IDF.

2.2.1.3 *Cosine Similarity*

Cosine similarity merupakan metode pengukuran kesamaan antar dua vektor dalam ruang multidimensi yang digunakan secara luas pada analisis teks. Dalam konteks *text mining*, vektor tersebut biasanya berasal dari representasi numerik kata atau dokumen yang diperoleh menggunakan metode TF-IDF. Menurut Widianto et al. (2024), *cosine similarity* digunakan untuk menentukan seberapa mirip dua dokumen berdasarkan sudut kosinus antara vektor-vektor penyusunnya semakin kecil sudut yang terbentuk, semakin besar tingkat kesamaannya.

Secara matematis, rumus dasar *cosine similarity* dapat dinyatakan sebagai berikut:

$$Cosine\ similarity(A, B) = \frac{A \cdot B}{\| A \| \times \| B \|} \qquad (2.\ 2)$$

Keterangan:
A = vektor dokumen pertama.
B = vektor dokumen kedua.
|| A || = panjang (magnitudo) vektor A.
|| B || = panjang (magnitudo) vektor B.

Nilai hasil perhitungan berada pada rentang 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa kedua teks memiliki kemiripan tinggi, sedangkan nilai mendekati 0 menandakan perbedaan yang besar (Chamira, 2022).

Dalam penelitian ini, nilai *cosine similarity* digunakan untuk mengukur tingkat kesesuaian antara judul dan deskripsi data budaya pada situs budaya-indonesia.org. Semakin tinggi nilai yang diperoleh, semakin kuat hubungan kesesuaian antar teks tersebut. Untuk mempermudah interpretasi hasil, tingkat kesesuaian dikategorikan pada Tabel 2.2 sebagai berikut:

Tabel 2. 2 Kategori Nilai *Cosine similarity*

Rentang Nilai *Cosine similarity*	**Kategori Kesesuaian**	**Interpretasi**
≥ 0,70	Sangat Sesuai	Judul dan deskripsi memiliki indikasi hubungan semantik yang kuat dan saling mendukung.
0,20 – 0,69	Cukup Sesuai	Terdapat keterkaitan sebagian antara isi deskripsi dan makna judul.
< 0,20	Tidak Sesuai	Deskripsi tidak memiliki indikasi keterhubungan semantik yang jelas dengan judul.

Pendekatan ini sejalan dengan Julians et al. (2024) yang menekankan pentingnya pengukuran kesesuaian teks berbasis *cosine similarity* untuk menilai relevansi antar teks dalam konteks budaya digital. Dengan demikian, penggunaan metode ini tidak hanya memberikan dasar kuantitatif untuk menilai hubungan antar elemen teks, tetapi juga memperkuat evaluasi terhadap konsistensi dan kualitas data budaya digital di Indonesia.

2.2.1.3.1 Penentuan Ambang *(Threshold) Cosine similarity*

Penentuan ambang *(threshold) cosine similarity* dalam penelitian ini dilakukan secara empiris dengan mempertimbangkan karakteristik data serta distribusi nilai *similarity* yang dihasilkan dari pasangan judul dan deskripsi pada dataset budaya-indonesia.org. *Cosine similarity* menghasilkan nilai kontinu dalam rentang 0 hingga 1, sehingga diperlukan pengelompokan nilai untuk mempermudah interpretasi tingkat kesesuaian teks secara sistematis.

Penelitian ini mengacu pada pendekatan yang digunakan oleh Bashir et al. (2021) yang tidak hanya menggunakan satu ambang batas, tetapi menetapkan dua threshold untuk *cosine similarity*, yaitu COS_LOWER = 0.2 dan COS_UPPER = 0.7. Dalam penelitian tersebut, nilai ≥ 0.2 diinterpretasikan sebagai teks yang

“cukup mirip secara semantik”, sedangkan nilai ≥ 0.7 menunjukkan teks yang “sangat mirip secara semantik” (Bashir et al., 2021). Penggunaan lebih dari satu threshold ini penting karena memungkinkan interpretasi kemiripan teks menjadi bertingkat, tidak sekadar “mirip” atau “tidak mirip”. Dengan demikian, hasil analisis dapat membedakan apakah keterkaitan judul dan deskripsi hanya lemah, sedang, atau sudah kuat.

Dengan penggunaan lebih dari satu threshold, hasil pengukuran *cosine similarity* tidak hanya digunakan untuk menentukan apakah suatu pasangan judul dan deskripsi mirip atau tidak, tetapi juga untuk mengelompokkan tingkat kesesuaiannya. Pendekatan ini memungkinkan proses evaluasi kesesuaian judul dan deskripsi dilakukan secara lebih objektif, terstruktur, dan mudah diinterpretasikan dalam konteks evaluasi kualitas data budaya digital. Oleh karena itu, penelitian ini menetapkan kategori *threshold* sebagai berikut:

1. **Nilai *cosine similarity* ≥ 0,70** dikategorikan sebagai *Sangat Sesuai*, yang menunjukkan adanya keterkaitan konten teks yang kuat antara judul dan deskripsi.
2. **Nilai *cosine similarity* antara 0,20 hingga 0,69** dikategorikan sebagai *Cukup Sesuai*, yang merepresentasikan keterkaitan parsial atau relevansi konteks sebagian.
3. **Nilai *cosine similarity* < 0,20** dikategorikan sebagai *Tidak Sesuai*, yang mengindikasikan minimnya hubungan konten teks antara judul dan deskripsi.

Pengelompokan nilai *cosine similarity* ini digunakan sebagai alat bantu interpretasi hasil analisis dan tidak dimaksudkan untuk mengukur kesesuaian semantik secara absolut. Dengan demikian, *threshold* yang digunakan berfungsi sebagai indikator tingkat kesesuaian teks berbasis representasi term, yang selanjutnya diperkuat melalui penerapan sistem berbasis aturan *(rule-based system)* pada tahap evaluasi akhir.

2.2.2 CRISP-DM *Framework*

CRISP-DM *(Cross-Industry Standard Process for Data Mining)* merupakan kerangka kerja yang banyak digunakan dalam penelitian berbasis *data mining* dan *text mining* karena menyediakan alur kerja yang sistematis dari perencanaan hingga evaluasi hasil (Plotnikova et al., 2022). Menurut Sudar et al., (2024), CRISP-DM membantu peneliti memahami masalah, menyiapkan data, membangun model, dan mengevaluasi hasilnya secara terstruktur sehingga hasil analisis menjadi lebih akurat dan replikatif.

Kerangka kerja ini terdiri dari lima tahapan utama yang bersifat iteratif, yaitu:

1 *Business Understanding*

Tahapan ini berfokus pada pemahaman tujuan, ruang lingkup, dan kebutuhan proyek analisis data. Pada fase ini ditentukan permasalahan inti yang ingin diselesaikan, sebagaimana dijelaskan dalam kerangka CRISP-DM bahwa pemahaman konteks masalah merupakan fondasi awal proses *data mining* (Elkabalawy et al., 2024).

2 *Data Understanding*

Fase ini mencakup pengumpulan data, eksplorasi awal, identifikasi struktur, serta evaluasi kualitas data. Pada tahapan ini dilakukan pemeriksaan kelengkapan, konsistensi, dan karakteristik awal data sebelum masuk ke proses persiapan. Eksplorasi awal ini penting untuk memahami potensi pola dan masalah pada dataset (Elkabalawy et al., 2024).

3 *Data Preparation*

Tahapan ini melibatkan pembersihan dan transformasi data agar siap dianalisis. Proses umum mencakup *cleaning*, *integration*, *transformation*, serta penyesuaian format data agar kompatibel dengan metode analisis yang akan digunakan (Elkabalawy et al., 2024).

4 *Modeling*

Pada fase ini, teknik atau algoritma dipilih dan diterapkan untuk menghasilkan model analisis data. Kerangka CRISP-DM menekankan bahwa beberapa pendekatan dapat diuji untuk menemukan model yang paling sesuai dengan karakteristik data dan tujuan penelitian (Elkabalawy et al., 2024).

5 *Evaluation*

Tahap ini bertujuan menilai hasil model dan memastikan bahwa model telah menjawab tujuan yang dirumuskan pada awal proses. Evaluasi dilakukan untuk menilai kualitas hasil, mengidentifikasi kekurangan model, serta menentukan kelayakan hasil untuk diimplementasikan (Elkabalawy et al., 2024).

Kerangka CRISP-DM dipilih karena sifatnya yang fleksibel dan sistematis, sehingga cocok diterapkan pada penelitian *text mining* yang memerlukan alur kerja terstruktur dari pemahaman masalah hingga evaluasi hasil. Pendekatan ini memastikan bahwa seluruh tahapan mulai dari perencanaan, pengumpulan, pengolahan, hingga analisis data teks budaya dilakukan secara terukur untuk menghasilkan temuan yang valid dan dapat dipertanggungjawabkan.

2.2.3 *Rule-Based System* dalam Analisis Teks

Sistem berbasis aturan (*rule-based system*) merupakan salah satu pendekatan dalam bidang *Natural Language Processing* (NLP) yang memanfaatkan seperangkat aturan linguistik eksplisit untuk menganalisis, mengelompokkan, atau memvalidasi teks (Pulungan, 2024). Tidak seperti metode statistik murni yang mengandalkan frekuensi dan bobot kata seperti TF-IDF, pendekatan ini meniru logika manusia dalam memahami bahasa dengan mendefinisikan hubungan kata, pola kalimat, dan makna kontekstual melalui pernyataan *if–then*. Menurut penelitian terbaru, sistem berbasis aturan kembali banyak digunakan dalam pendekatan *hybrid NLP* karena mampu meningkatkan sensitivitas indikasi semantik, terutama pada data berbahasa alami yang memiliki ragam lokal seperti Bahasa Indonesia (Widianto et al., 2024).

Dalam konteks penelitian ini, sistem rule-based digunakan karena metode statistik berbasis pembobotan kata belum sepenuhnya mampu menangkap konteks makna dan pola linguistik eksplisit pada teks deskriptif berbahasa Indonesia. Keterbatasan ini juga diidentifikasi dalam penelitian Aubaid et al., (2024), yang menunjukkan bahwa metode statistik seperti Naive Bayes dan cosine similarity

memiliki keterbatasan dalam menangkap hubungan semantik domain-spesifik, sehingga memerlukan dukungan aturan eksplisit berbasis keyword untuk meningkatkan akurasi dan transparansi keputusan klasifikasi. Misalnya, entri budaya dengan judul "onde-onde" dapat memiliki deskripsi yang relevan, seperti "kue tradisional dari tepung ketan berisi kacang hijau," namun sistem TF-IDF murni akan menganggap keduanya tidak mirip karena tidak menemukan kata "onde-onde" di deskripsi. Dengan demikian, pendekatan ini sejalan dengan kerangka kerja W2vRule yang dikembangkan oleh Aubaid et al., (2024), di mana rules dirancang berdasarkan keyword domain-spesifik yang ditemukan dalam dataset dan bekerja sebagai lapisan keputusan yang melengkapi perhitungan similarity — tanpa mengubah nilai skor statistik yang telah dihitung. Integrasi tersebut terbukti meningkatkan akurasi klasifikasi hingga 89.91% pada dataset Reuters-21578, yang secara metodologis sebanding dengan kebutuhan evaluasi kesesuaian teks budaya dalam penelitian ini di mana rules dirancang berdasarkan pola linguistik yang ditemukan pada dataset budaya Indonesia.

Aturan dalam sistem ini disusun berdasarkan pola umum teks dan dirancang untuk menilai kualitas deskripsi secara logis. Jenis-jenis aturan yang digunakan antara lain sebagai berikut:

1. ***Rule* Deskripsi Pendek**

   yaitu aturan yang menandai entri dengan panjang deskripsi kurang dari lima kata sebagai Tidak Valid. Aturan ini berfungsi untuk memfilter data yang tidak cukup informatif dan berpotensi menurunkan akurasi perhitungan.

2. ***Rule* Keterkaitan Kata Judul**

   yaitu aturan yang mengevaluasi seberapa banyak kata dari judul muncul di deskripsi. Jika satu hingga tiga kata dari judul muncul di deskripsi, entri tersebut dikategorikan *Cukup Valid* karena menunjukkan adanya kesesuaian dasar antara keduanya.

3. ***Rule* Konteks Budaya**

   yang digunakan untuk mendeteksi relevansi teks berdasarkan kategori budaya seperti tarian, makanan, alat musik, atau upacara. Misalnya, keberadaan pola frasa seperti "terbuat dari," "digunakan pada upacara," atau "alat musik tradisional" akan menaikkan bobot kesesuaian karena menandakan hubungan langsung antara deskripsi dan kategori budaya yang sesuai.

4. ***Rule* Sinonim dan Variasi Daerah**

   yang berfungsi untuk mengenali istilah dengan makna sepadan atau penggunaan kata khas daerah. Contohnya, kata *"tari"* dapat disetarakan dengan *"menari"*, *"nyanyian"* dengan *"lagu"*, atau *"gamelan Jawa"* dengan *"alat musik tradisional"*. Penerapan aturan ini memungkinkan sistem untuk memahami makna lokal yang tidak terdeteksi oleh TF-IDF, sekaligus memperkaya analisis indikasi semantik.

5. ***Rule* Pola Frasa Deskriptif**

   yaitu aturan yang mendeteksi pola frasa yang sering digunakan dalam menjelaskan fungsi, bahan, atau konteks sosial dari artefak budaya. Beberapa frasa yang sering muncul antara lain "bahan utama," "digunakan

oleh masyarakat," dan "memiliki makna simbolik." Pola-pola ini menjadi indikator bahwa deskripsi memiliki konteks budaya yang kuat, meskipun tidak menyebut judul secara langsung.

6. ***Rule* Anti-Spam Pengulangan**

   yaitu aturan untuk menghapus pengaruh pengulangan kata atau frasa yang sama sebanyak tiga kali atau lebih secara berurutan, seperti "onde-onde onde-onde onde-onde" atau "sate sate sate." Pengulangan seperti ini sering muncul akibat kesalahan input atau duplikasi otomatis yang dapat menghasilkan skor TF-IDF tinggi secara tidak wajar. Kata atau frasa yang terdeteksi dihapus dari proses perhitungan agar bobot kesamaan tidak menjadi bias.

7. ***Rule* Daftar Bahan Tanpa Konteks**

   yang mendeteksi deskripsi yang hanya berisi daftar bahan atau instruksi memasak tanpa keterangan makna budaya. Misalnya, jika deskripsi hanya berisi kata seperti "tepung ketan, kacang hijau, wijen, minyak goreng" tanpa penjelasan simbolik atau fungsi sosial, sistem menurunkan bobot kesesuaian karena teks tersebut lebih bersifat teknis daripada informatif secara budaya.

Setiap aturan di atas bekerja secara berurutan. Struktur kerja berlapis ini secara konseptual sejalan dengan pendekatan *rule-based classification* yang diterapkan oleh Aubaid et al., (2024), di mana rules dirancang berdasarkan *keyword* dan pola domain-spesifik yang ditemukan dalam dataset, kemudian ditempatkan sebagai lapisan keputusan yang bekerja secara independen dari perhitungan

similarity sehingga nilai *cosine similarity* tidak diubah, melainkan hasil keputusan kategori yang disesuaikan. Sistem pertama-tama memeriksa panjang deskripsi, kemudian menghapus kata berulang, mengevaluasi keterkaitan kata judul, dan mendeteksi keberadaan sinonim serta pola frasa budaya. Hasil evaluasi ini kemudian digunakan untuk menyesuaikan kategori kesesuaian yang diperoleh dari perhitungan *cosine similarity* TF-IDF, tanpa mengubah nilai skornya. Misalnya, jika skor *cosine* berada pada kisaran menengah (0,20–0,69) namun deskripsi memenuhi aturan konteks budaya dan sinonim, maka kategori akhir dapat dinaikkan menjadi Sangat Sesuai. Sebaliknya, jika deskripsi hanya berupa daftar bahan atau teks sangat pendek, maka kategori akhir diturunkan menjadi Tidak Sesuai meskipun skor TF-IDF awalnya berada pada kisaran menengah.

Pendekatan ini sejalan dengan penelitian oleh Widianto et al. (2024) yang menunjukkan bahwa kombinasi antara metode statistik dan sistem berbasis aturan dapat meningkatkan keakuratan analisis teks hingga 20% dibandingkan metode tunggal. Selain itu, sistem *rule-based* menawarkan transparansi dan interpretabilitas yang tinggi, karena setiap keputusan dapat dijelaskan melalui aturan yang eksplisit. Hal ini sangat penting untuk data budaya yang memiliki variasi bahasa, istilah lokal, dan perbedaan struktur narasi antar daerah. Dengan demikian, penggunaan *rule-based system* dalam penelitian ini tidak hanya meningkatkan ketepatan pengukuran kesesuaian teks antara judul dan deskripsi, tetapi juga memperkuat indikasi validitas semantik dalam konteks pelestarian data budaya digital Indonesia.

2.3 Sinopsis

Penelitian ini berfokus pada pengukuran kesesuaian antara judul dan deskripsi data budaya yang bersumber dari situs budaya-indonesia.org. Platform ini merupakan repositori digital terbuka hasil kontribusi publik (*crowdsourcing*) yang menyimpan puluhan ribu entri budaya seperti tarian, makanan, dan alat musik tradisional. Permasalahan yang muncul ialah banyaknya entri dengan deskripsi tidak relevan, terlalu singkat, atau hanya berisi tautan sumber. Untuk itu digunakan pendekatan *text mining* berbasis *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *cosine similarity*, sebagaimana diterapkan secara efektif oleh (Amrulloh & Adam, 2021; Widianto et al., 2024) dalam pengukuran kemiripan teks berbahasa Indonesia.

Selain pendekatan statistik tersebut, penelitian ini mencoba memperkuat model dengan *rule-based system* yang mendeteksi pola linguistik eksplisit pada deskripsi budaya. Berdasarkan konsep Opitz et al. (2025), aturan linguistik mampu memperbaiki hasil analisis teks melalui identifikasi deskripsi pendek, kemunculan kata dari judul, sinonim budaya (*tari* ≈ *menari*, *lagu* ≈ *nyanyian*), serta pengulangan berlebihan yang menurunkan validitas makna.

Seluruh proses analisis dirancang mengikuti kerangka CRISP-DM, mencakup tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, dan *evaluation*, untuk memastikan penelitian berjalan sistematis dan dapat direplikasi. Pendekatan hibrida ini diharapkan dapat menghasilkan model evaluasi semantik yang andal, transparan, dan relevan bagi peningkatan kualitas data budaya digital di Indonesia.

BAB III

DESAIN DAN IMPLEMENTASI

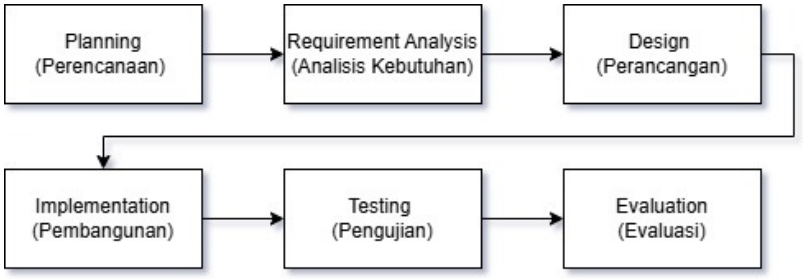
3.1 Desain Penelitian

Penelitian ini merupakan penelitian deskriptif kuantitatif yang menggunakan pendekatan *text mining* untuk mengevaluasi tingkat kesesuaian antara kolom judul dan deskripsi pada dataset budaya dari situs budaya-indonesia.org. Pendekatan ini dipilih karena *text mining* terbukti efektif dalam mengekstraksi pengetahuan dari teks dan mendeteksi indikasi hubungan semantik antar dokumen (Julians et al., 2024).

Metode utama yang digunakan adalah *Term Frequency–Inverse Document Frequency* (TF-IDF) yang berfungsi untuk menghitung bobot kemunculan kata dalam dokumen. TF-IDF telah banyak diterapkan dalam pengukuran kesamaan teks dan validasi data berbasis teks di berbagai penelitian (Amrulloh & Adam, 2021; Wibowo & Salim, 2024). Dalam penelitian ini, metode tersebut dipadukan dengan pendekatan *rule-based system* sederhana untuk menangkap pola linguistik seperti kesamaan sinonim dan struktur kalimat yang tidak terdeteksi oleh pendekatan statistik (Pulungan, 2024).

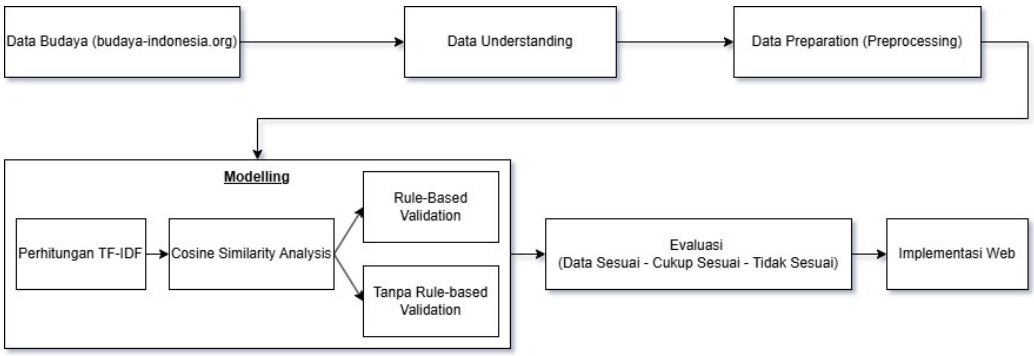
Dalam konteks pengembangan sistem analisis, penelitian ini disusun menggunakan pendekatan *System Development Life Cycle* (SDLC) dengan model *Waterfall*. Model ini dipilih karena bersifat terstruktur, berurutan, dan sesuai untuk proses pengembangan aplikasi web yang dimulai dari identifikasi kebutuhan hingga evaluasi hasil akhir. Tahapan SDLC yang diterapkan meliputi: (1) *Perencanaan*,

yaitu penentuan tujuan evaluasi data budaya; (2) *Analisis*, yaitu identifikasi permasalahan ketidaksesuaian semantik akibat sifat kontribusi *crowdsourced;* (3) *Perancangan*, yaitu penyusunan alur pemodelan TF-IDF, *cosine similarity*, dan *rule-based system*; (4) *Implementasi*, yaitu pembangunan sistem analisis menggunakan bahasa pemrograman Python; (5) *Pengujian*, yaitu verifikasi hasil pada sampel data; dan (6) *Evaluasi*, yaitu penilaian konsistensi dan ketepatan hasil klasifikasi. Ditunjukkan dalam Gambar 3.1 berikut.



Gambar 3. 1 Bagan Desain SDLC

Adapun kerangka kerja CRISP-DM *(Cross Industry Standard Process for Data Mining)* tetap digunakan pada bagian pemodelan analisis teks untuk mengatur urutan pemrosesan data, yang meliputi tahapan *Business Understanding, Data Understanding, Data Preparation, Modeling,* dan *Evaluation* (Sari et al., 2021). Berikut merupakan kerangka CRISP-DM dalam Gambar 3.2.



Gambar 3. 2 Bagan Desain CRISP-DM

Tujuan dari desain penelitian dan desain sistem ini adalah untuk mengukur tingkat kesesuaian antara judul dan deskripsi data budaya secara kuantitatif serta mengidentifikasi pola ketidaksesuaian teks yang dapat memengaruhi kualitas informasi budaya digital. SDLC digunakan untuk menggambarkan alur pengembangan sistem secara umum, sedangkan CRISP-DM digunakan secara spesifik sebagai kerangka kerja analisis data dan *text mining*. Kedua pendekatan ini tidak tumpang tindih, melainkan saling melengkapi pada level yang berbeda. Hasil penelitian ini diharapkan dapat menjadi dasar bagi pengelola platform budaya digital untuk meningkatkan akurasi, relevansi, dan kredibilitas data budaya Indonesia.

3.2 Sumber Data

Penelitian ini menggunakan data yang bersumber dari situs budaya-indonesia.org, yang merupakan domain resmi proyek Perpustakaan Digital Budaya Indonesia (PDBI) yang dikelola oleh *Indonesian Archipelago Culture Initiatives (IACI)*.

Platform ini berfungsi sebagai repositori terbuka artefak budaya Indonesia hasil kontribusi publik (*crowdsourcing*). PDBI menghimpun beragam bentuk warisan budaya mulai dari tarian, musik, kain tradisional, makanan khas, ornamen, cerita rakyat, hingga ritual dan senjata tradisional dengan tujuan menjadi referensi daring nasional yang terdokumentasi dan dapat diakses secara luas.

Situs ini dirancang agar masyarakat dapat menambahkan data budaya daerahnya masing-masing. Karena sifatnya yang *crowdsourced*, tingkat kelengkapan dan kualitas informasi tiap entri sangat bervariasi: beberapa deskripsi memuat narasi budaya yang kaya, sedangkan lainnya hanya berisi tautan atau frasa pendek. Variasi kualitas data inilah yang menjadi latar penting penelitian ini, karena relevansi antara judul dan deskripsi perlu diverifikasi agar informasi budaya yang disajikan tetap akurat dan kredibel.

Data yang digunakan dalam penelitian berjumlah sekitar ± 58 000 entri yang dapat diakses publik pada situs budaya-indonesia.org. Pengambilan data dilakukan dengan cara mengekspor secara langsung di web budaya-indonesia.org dalam bentuk csv dan excel. Analisis difokuskan pada dua atribut utama berbasis teks, yaitu:

Tabel 3. 1 Atribut Utama

Atribut	**Tipe**	**Keterangan Penggunaan**
Judul	Teks	Nama artefak budaya; digunakan sebagai acuan utama dalam pengukuran kesesuaian.
Deskripsi	Teks	Uraian singkat artefak; dianalisis menggunakan TF-IDF dan divalidasi melalui *rule-based system*.

Hanya entri dengan teks deskripsi nonkosong yang diikutsertakan. Entri dengan deskripsi yang berisi tautan eksternal saja atau media nonteks (gambar, video, atau dokumen) dikecualikan dari analisis.

3.3 Tahapan Penelitian

Tahapan penelitian ini disusun berdasarkan kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri dari enam fase utama, namun dalam penelitian ini difokuskan pada empat tahap inti yang relevan dengan proses analisis teks, yaitu *Business & Data Understanding*, *Data Preparation*, *Modelling*, serta *Evaluation & Visualization*. Setiap tahap saling berkaitan dan membentuk alur sistematis mulai dari pemahaman konteks serta karakteristik data, pengolahan dan pembersihan teks, pembangunan model analisis kesesuaian, hingga evaluasi dan visualisasi hasil. Pendekatan ini digunakan untuk memastikan bahwa proses analisis dilakukan secara terstruktur, terukur, dan dapat direplikasi sesuai tujuan penelitian.

3.3.1 Business & Data Understanding

Tahapan *Business Understanding* berfokus pada pemahaman konteks dan tujuan utama penelitian, yaitu mengevaluasi kesesuaian antara kolom *judul* dan *deskripsi* pada data budaya di situs budaya-indonesia.org (PDBI). Tujuan utama tahap ini adalah untuk mengidentifikasi permasalahan kualitas data budaya digital yang muncul akibat kontribusi publik (*crowdsourcing*), seperti ketidaksesuaian konteks, pengulangan informasi, dan deskripsi yang tidak relevan.

Menurut Plotnikova et al. (2022), tahap awal pada kerangka CRISP-DM menuntut pemahaman mendalam terhadap tujuan dan konteks analisis agar hasil *data mining* dapat menyelesaikan masalah nyata secara terarah dan terukur. Dalam konteks penelitian ini, permasalahan yang dihadapi adalah ketidaksesuaian makna

antara judul dan deskripsi dalam entri budaya yang mengakibatkan menurunnya akurasi representasi pengetahuan budaya digital.

Tahapan *Data Understanding* dilakukan untuk mengenali struktur dan karakteristik data teks yang dianalisis, seperti panjang deskripsi, keberagaman istilah daerah, serta frekuensi kemunculan kata. Menurut Julians et al. (2024), analisis konten budaya kolaboratif memerlukan eksplorasi menyeluruh terhadap data teks hasil kontribusi masyarakat, karena variasi linguistik dan ketidakteraturan struktur kalimat dapat memengaruhi hasil pengolahan. Pemahaman konteks ini penting untuk menilai kualitas data serta menentukan strategi pembersihan dan transformasi teks sebelum tahap pemodelan dilakukan.

Selanjutnya, Pulungan (2024) menjelaskan bahwa dalam tahap pemahaman data, perlu ditentukan indikator kesesuaian semantik yang menjadi dasar penilaian hubungan makna antar teks. Dalam penelitian ini, indikator tersebut mencakup:

1. Keterkaitan kata utama antara judul dan deskripsi, yang diukur menggunakan *cosine similarity*.
2. Konteks budaya yang teridentifikasi dalam deskripsi, seperti istilah khas untuk kategori artefak budaya (tarian, makanan, alat musik, atau upacara).
3. Kelengkapan dan panjang deskripsi, sebagai dasar untuk menilai kecukupan indikasi konteks semantik.
4. Kesepadanan sinonim dan istilah lokal, yang ditangani menggunakan *rule-based system* agar variasi bahasa daerah tetap terakomodasi.

Pendekatan tersebut sejalan dengan pandangan Sudar et al. (2024) yang menegaskan bahwa dalam penerapan CRISP-DM untuk analisis teks publik,

pemahaman terhadap konteks sosial dan karakteristik data menjadi fondasi utama sebelum model dibangun. Dengan demikian, tahap *Business & Data Understanding* pada penelitian ini memastikan bahwa analisis kesesuaian dilakukan berdasarkan pemahaman yang komprehensif terhadap konteks budaya, struktur teks, dan kebutuhan evaluasi kualitas data digital Indonesia.

3.3.2 Data Preparation

Tahap *Data Preparation* merupakan proses penting dalam kerangka kerja CRISP-DM, di mana data mentah diubah menjadi format bersih dan siap dianalisis. Menurut Plotnikova et al. (2022), tahap ini menentukan keberhasilan model analisis karena hasil yang baik hanya dapat diperoleh dari data yang terstruktur dan bebas *noise*. Dalam konteks penelitian ini, tahap *preprocessing* diterapkan pada dua kolom teks utama dari situs budaya-indonesia.org, yaitu judul dan deskripsi, untuk memastikan keseragaman format sebelum dilakukan perhitungan TF-IDF dan *cosine similarity*.

Data yang dikumpulkan dari situs tersebut sangat beragam dalam gaya penulisan, panjang teks, serta penggunaan istilah daerah. Oleh karena itu, dilakukan serangkaian tahapan *text cleaning* dan *normalization* agar teks memiliki struktur linguistik yang seragam. Tahapan ini meliputi:

1. *Case Folding*

   Semua huruf diubah menjadi huruf kecil (*lowercase*) agar sistem tidak membedakan antara huruf kapital dan huruf kecil. Seperti contoh pada Tabel 3.2 berikut.

Tabel 3. 2 Contoh Proses *Case Folding*

Kolom	**Sebelum**	**Sesudah**
Judul	Tari Kecak Bali	tari kecak bali
Deskripsi	Tari Kecak merupakan tarian tradisional yang berasal dari Bali	tari kecak merupakan tarian tradisional yang berasal dari bali

Proses pada Tabel 3.2 diatas mengurangi redundansi kata serta meningkatkan konsistensi analisis (Amrulloh & Adam, 2021).

2. *Tokenizing*

Teks dipecah menjadi potongan kata atau token menggunakan pustaka NLTK. Tahapan ini memungkinkan sistem menghitung frekuensi kata yang menjadi dasar pembobotan TF-IDF. Seperti contoh pada Tabel 3.3 berikut.

Tabel 3. 3 Contoh *Proses Tokenizing*

Kolom	**Hasil Tokenisasi**
Judul	[tari, kecak, bali]
Deskripsi	[tari, kecak, merupakan, tarian, tradisional, yang, berasal, dari, bali]

3. *Stopword Removal*

Kata-kata umum seperti *"yang"*, *"dan"*, atau *"dari"* dihapus menggunakan daftar *stopword* Bahasa Indonesia dari pustaka NLTK dan Sastrawi. Seperti contoh pada Tabel 3.4 berikut.

Tabel 3. 4 Contoh Proses *Stopword Removal*

Kolom	**Sebelum**	**Sesudah**
Judul	tari kecak bali	tari kecak bali
Deskripsi	tari kecak merupakan tarian tradisional yang berasal dari bali	tari kecak tarian tradisional berasal bali

Menurut Widianto et al. (2024), penghapusan kata tidak penting dapat meningkatkan fokus analisis terhadap kata bermakna utama dan mengurangi *noise*.

4. *Stemming*

   Setiap kata diubah ke bentuk dasarnya menggunakan pustaka Sastrawi. Misalnya, kata *"penari"*, *"tarian"*, dan *"menari"* menjadi *"tari"*. Seperti contoh pada Tabel 3.5 berikut.

Tabel 3. 5 Contoh Proses *Stemming*

Kolom	**Sebelum**	**Sesudah**
Judul	tari kecak bali	tari kecak bali
Deskripsi	tari kecak tarian tradisional berasal bali	tari kecak tari tradisional asal bali

   Menurut Pulungan (2024), *stemming* berfungsi menjaga keseragaman makna dan meminimalkan perbedaan bentuk kata yang sebenarnya memiliki akar semantik sama.

5. *Cleaning Noise*

   Karakter khusus, angka, dan tanda baca dihapus menggunakan *Regular Expression (Regex)*. Tahap ini memastikan teks bersih dari elemen non-linguistik yang dapat mengganggu perhitungan kemiripan (Sudar et al., 2024).

Seluruh tahapan dilakukan secara berurutan menggunakan bahasa pemrograman Python dengan kombinasi pustaka NLTK, Sastrawi, dan Regex. Hasil akhirnya berupa teks yang telah distandarkan, terlepas dari variasi ejaan dan gaya penulisan, serta siap untuk tahap Modeling. Berikut merupakan contoh hasil Preprocessing pada Tabel 3.6

Tabel 3. 6 Contoh Hasil *Preprocessing*

Sebelum Preprocessing	**Hasil Akhir**
Tari Kecak Bali	tari kecak bali
Tari Kecak merupakan tarian tradisional yang berasal dari Bali.	tari kecak tari tradisional asal bali

Pendekatan ini sejalan dengan Julians et al. (2024) yang menegaskan bahwa proses *text preprocessing* menjadi fondasi penting dalam menjaga keakuratan analisis kesesuaian dan validitas data budaya digital Indonesia.

3.3.3 *Modelling*

Tahapan *modeling* merupakan inti dari penelitian ini, yaitu proses membangun model untuk mengukur tingkat kesesuaian antara kolom *judul* dan *deskripsi* pada data budaya di situs budaya-indonesia.org. Pemodelan ini dilakukan dengan tiga pendekatan utama: pembobotan kata menggunakan metode TF-IDF, pengukuran kemiripan menggunakan *cosine similarity*, dan penyempurnaan hasil dengan sistem berbasis aturan (*rule-based system*).

3.3.3.1 Perhitungan TF-IDF

Tahap pertama adalah representasi teks ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Sari et al. (2021) dalam *Jurnal Terapan Informatika Nusantara* menjelaskan bahwa TF-IDF mampu memberikan bobot berbeda bagi setiap kata berdasarkan tingkat kepentingannya di dalam suatu dokumen dan keseluruhan korpus. Pendekatan ini terbukti meningkatkan akurasi klasifikasi dokumen hingga 98%. Dalam implementasinya, pembentukan model TF-IDF dilakukan dengan menggabungkan seluruh teks judul dan deskripsi ke dalam satu korpus. Penggabungan ini bertujuan agar *vocabulary* yang terbentuk bersifat konsisten dan kedua jenis teks berada dalam ruang fitur yang sama. Dengan demikian, vektor judul dan vektor deskripsi dapat dibandingkan secara langsung tanpa perbedaan dimensi fitur. Proses

pembobotan dilakukan menggunakan algoritma *TfidfVectorizer* dengan konfigurasi ngram_range = (1,2), sehingga sistem mempertimbangkan unigram dan bigram.

Hasil dari proses ini adalah sebuah matriks TF-IDF berdimensi $(2n \times m)$, dengan n sebagai jumlah data dan m sebagai jumlah fitur yang terbentuk. Matriks tersebut kemudian dipisahkan kembali menjadi dua bagian, yaitu matriks vektor judul $(n \times m)$ dan matriks vektor deskripsi $(n \times m)$, sehingga setiap entri memiliki dua representasi vektor dalam ruang fitur identik. Sebagai ilustrasi, berikut disajikan contoh perhitungan manual TF-IDF pada korpus sederhana yang merepresentasikan proses yang sama dalam sistem.

a. Pembentukan Korpus dan Vocabulary

   Korpus dibentuk dari gabungan seluruh teks judul (D1–D3) dan deskripsi (D4–D6), sehingga jumlah dokumen N = 6. Dari keenam dokumen tersebut diekstrak vocabulary yang terdiri dari m = 15 fitur unik, seperti yang bisa dilihat pada Tabel 3.7 berikut.

Tabel 3. 7 Korpus Gabungan Judul dan Deskripsi

Dokumen	**Teks**
D1	tari kecak bali
D2	angklung jawa barat
D3	rendang sumatera barat
D4	tari kecak tari tradisional asal bali
D5	alat musik bambu tradisional jawa barat
D6	makanan khas tradisional sumatera barat

b. Perhitungan *Term Frequency (TF)*

   Untuk membandingkan judul dan deskripsi pada pasangan yang sama, contoh perhitungan difokuskan pada D1 (“Tari Kecak Bali”) sebagai judul

dan D4 (“Tari Kecak Tari Tradisional Asal Bali”) sebagai deskripsinya. TF menghitung frekuensi kemunculan setiap term dalam dokumen tersebut.

TF dihitung menggunakan rumus $TF(t,d) = f(t,d) \,/\, \Sigma k\, f(k,d)$

dengan f(t,d) adalah frekuensi kemunculan term t dalam dokumen d, dan Σk f(k,d) adalah total seluruh term dalam dokumen d. Normalisasi ini memastikan dokumen dengan panjang berbeda dapat dibandingkan secara adil. Untuk D1 total kata = 3, sedangkan D4 total kata = 6. Hasil perhitungannya akan ditampilkan pada Tabel 3.8 dan Tabel 3.9

Tabel 3. 8 Frekuensi Kemunculan Term pada D1

Term	**frekuensi**	**TF = f / Σf (total = 3)**
tari	1	1/3 = 0,333
kecak	1	1/3 = 0,333
bali	1	1/3 = 0,333
tradisional	0	0/3 = 0
asal	0	0/3 = 0

$$\boldsymbol{TF = f \,/\, \Sigma f\ (total = 3)}$$

$$tari = 1/3 = 0{,}333$$

$$kecak = 1/3 = 0{,}333$$

$$bali = 1/3 = 0{,}333$$

$$tradisional = 0/3 = 0$$

$$asal = 0/3 = 0$$

Tabel 3. 9 Frekuensi Kemunculan Term pada D4

Term	**Frekuensi**	**TF = f / Σf (total = 6)**
tari	2	2/6 = 0,333
kecak	1	1/6 = 0,167
bali	1	1/6 = 0,167
tradisional	1	1/6 = 0,167
asal	1	1/6 = 0,167

$$\boldsymbol{TF = f / \Sigma f\ (total = 6)}$$

$$tari = 2/6 = 0{,}333$$

$$kecak = 1/6 = 0{,}167$$

$$bali = 1/6 = 0{,}167$$

$$tradisional = 1/6 = 0{,}167$$

$$asal = 1/6 = 0{,}167$$

Perlu diperhatikan pada Tabel 3.9 bahwa meskipun kata tari muncul 2 kali di D4, nilai TF-nya tetap 0,333 karena dibagi total 6 kata. Hal ini berbeda dengan D1 pada Tabel 3.8 yang meskipun frekuensi mentah kata tari sama (1 kali), nilai TF-nya juga 0,333 karena total katanya hanya 3. Normalisasi ini mencegah bias akibat perbedaan panjang dokumen.

c. Perhitungan *Document Frequency (DF)* dan *Inverse Document Frequency (IDF)*

DF dihitung dari seluruh enam dokumen dalam korpus. IDF kemudian dihitung menggunakan rumus $IDF(t) = log10(N / df(t))$ dengan N = 6 (jumlah dokumen). Hasil perhitungan DF dan IDF untuk lima term kunci disajikan pada Tabel 3.10 berikut.

Tabel 3. 10 Nilai DF dan IDF per Term

Term	DF	N/DF	IDF = $\log_{10}$(N/DF)
tari	2	6/2 = 3	$\log_{10}(3) = 0{,}477$
kecak	2	6/2 = 3	$\log_{10}(3) = 0{,}477$
bali	2	6/2 = 3	$\log_{10}(3) = 0{,}477$
tradisional	3	6/3 = 2	$\log_{10}(2) = 0{,}301$
asal	1	6/1 = 6	$\log_{10}(6) = 0{,}778$

$tari : df = 2, IDF = log_{10}(6/2) = log_{10}(3) = 0{,}477$

$kecak : df = 2, IDF = log_{10}(6/2) = log_{10}(3) = 0{,}477$

$bali : df = 2, IDF = log_{10}(6/2) = log_{10}(3) = 0{,}477$

$tradisional : df = 3, IDF = log_{10}(6/3) = log_{10}(2) = 0{,}301$

$asal : df = 1, IDF = log_{10}(6/1) = log_{10}(6) = 0{,}778$

Dari Tabel 3.10 kata asal memperoleh nilai IDF tertinggi (0,778) karena hanya muncul pada satu dokumen sehingga dianggap paling spesifik. Sebaliknya, kata tradisional memiliki IDF lebih rendah (0,301) karena muncul pada tiga dokumen, sementara kata tari, kecak, dan bali memiliki IDF sedang (0,477) karena muncul pada dua dokumen.

d. Perhitungan Bobot TF-IDF

Bobot TF-IDF setiap term dihitung dengan rumus:

$TF - IDF(t, d) = TF(t, d)\ x\ IDF(t)$ yang bisa dilihat pada Tabel 3.11 berikut.

Tabel 3. 11 Bobot TF-IDF Judul D1

Term	TF	IDF	TF-IDF = TF × IDF
tari	0,333	0,477	0,159
kecak	0,333	0,477	0,159
bali	0,333	0,477	0,159
tradisional	0	0,301	0
asal	0	0,778	0

$tari : 0{,}333 \times 0{,}477 = 0{,}159$

$kecak : 0{,}333 \times 0{,}477 = 0{,}159$

$bali : 0{,}333 \times 0{,}477 = 0{,}159$

$tradisional : 0 \times 0{,}301 = 0$

$asal : 0 \times 0{,}778 = 0$

Dari Tabel 3.11 diperoleh vektor D1 (Judul): A = [0,159; 0,159; 0,159; 0; 0]. Berikut merupakan Tabel 3.12 untuk bobot TF-IDF deskripsi.

Tabel 3. 12 Bobot TF-IDF Deskripsi D4

Term	**TF**	**IDF**	**TF-IDF = TF × IDF**
tari	0,333	0,477	0,159
kecak	0,167	0,477	0,080
bali	0,167	0,477	0,080
tradisional	0,167	0,301	0,050
asal	0,167	0,778	0,130

$tari : 0{,}333 \times 0{,}477 = 0{,}159$

$kecak : 0{,}167 \times 0{,}477 = 0{,}080$

$bali : 0{,}167 \times 0{,}477 = 0{,}080$

$tradisional : 0{,}167 \times 0{,}301 = 0{,}050$

$asal : 0{,}167 \times 0{,}778 = 0{,}130$

Dari Tabel 3.12 diperoleh vektor D4 (Deskripsi): B = [0,159; 0,080; 0,080; 0,050; 0,130]. Perbedaan nilai antar term pada D4 mencerminkan kombinasi antara kelangkaan term di seluruh korpus (IDF) dan proporsi kemunculannya dalam dokumen (TF).

3.3.3.2 Pengukuran Kemiripan Menggunakan *Cosine Similarity*

Setelah diperoleh representasi vektor untuk judul dan deskripsi, tahap selanjutnya adalah menghitung tingkat kemiripan menggunakan metode *cosine similarity.* Widianto et al. (2024) dalam *Journal of Dinda* membuktikan bahwa kombinasi TF-IDF dan *cosine similarity* efektif dalam mengukur kemiripan indikasi semantik teks dengan tingkat akurasi hingga 93,6 %.

Dalam penelitian ini, perhitungan *cosine similarity* dilakukan dengan membentuk matriks kemiripan antara seluruh vektor judul dan seluruh vektor deskripsi. Namun, tujuan analisis bukan membandingkan seluruh dokumen secara silang, melainkan mengevaluasi kesesuaian judul dan deskripsi pada entri yang sama. Oleh karena itu, skor kesesuaian yang digunakan merupakan nilai diagonal dari matriks *cosine similarity*, yang merepresentasikan kemiripan antara pasangan judul dan deskripsi pada indeks yang identik. Pendekatan ini memastikan bahwa nilai similarity benar-benar mencerminkan konsistensi internal antara dua elemen teks dalam satu entri data budaya. Nilai hasil perhitungan berkisar antara 0 – 1, dengan interpretasi sebagai berikut:

≥ 0,70 = Sangat Sesuai; 0,20 – 0,69 = Cukup Sesuai; < 0,20 = Tidak Sesuai. Berikut adalah contoh perhitungan *cosine similarity* pada Tabel 3.13 menggunakan vektor D1 dan D4 yang telah diperoleh.

Tabel 3. 13 Perhitungan Dot Product dan Komponen Panjang Vektor D1

Tern	**A (D1)**	**B (D4)**	**A · B**	**A^2**	**B^2**
tari	0,159	0,159	0,159 × 0,159 = 0,025	$0{,}159^2 = 0{,}025$	$0{,}159^2$
kecak	0,159	0,080	0,159 × 0,080 = 0,013	$0{,}159^2 = 0{,}025$	$0{,}080^2$
bali	0,159	0,080	0,159 × 0,080 = 0,013	$0{,}159^2 = 0{,}025$	$0{,}080^2$
tradisional	0	0,050	0 × 0,050 = 0	$0^2 = 0$	$0{,}050^2$
asal	0	0,130	0 × 0,130 = 0	$0^2 = 0$	$0{,}130^2$
			0,051	**0,076**	**0,057**

Berdasarkan Tabel 3.13 di atas, diperoleh nilai dot product A·B = 0,051 Selanjutnya, panjang masing-masing vektor dihitung sebagai berikut.

Untuk vektor A (D1):

$$\| A \| = \sqrt{0,159^2 + 0,159^2 + 0,159^2 + 0^2 + 0^2}$$

$$\| A \| = \sqrt{0{,}025 + 0{,}025 + 0{,}025 + 0 + 0}$$

$$\| A \| = \sqrt{0{,}076} = 0{,}276$$

Untuk vektor B (D4):

$$\| B \| = \sqrt{0,159^2 + 0,080^2 + 0,080^2 + 0,050^2 + 0,130^2}$$

$$\| B \| = \sqrt{0{,}025 + 0{,}006 + 0{,}006 + 0{,}003 + 0{,}017}$$

$$\| B \| = \sqrt{0{,}057} = 0{,}239$$

Dengan demikian, nilai *cosine similarity* antara vektor A dan B dapat dihitung menggunakan persamaan:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\| A \| \times \| B \|}$$

$$= \frac{0{,}051}{0{,}276 \times 0{,}239} = \frac{0{,}051}{0{,}066} = 0{,}773$$

Nilai cosine similarity sebesar 0,773 menunjukkan bahwa pasangan D1 dan D4 termasuk dalam kategori "Sangat Sesuai" (≥ 0,70), yang berarti deskripsi "Tari Kecak Tari Tradisional Asal Bali" secara konsisten mencerminkan judul "Tari Kecak Bali". Nilai ini diperoleh karena kedua teks berbagi term-term kunci yang sama (tari, kecak, bali) dengan bobot TF-IDF yang signifikan, sehingga sudut antara kedua vektor relatif kecil.

3.3.3.3 Validasi Menggunakan *Rule-Based System*

Tahapan selanjutnya adalah penerapan *rule-based system* sebagai mekanisme validasi untuk melengkapi hasil pengukuran kesesuaian berbasis TF-IDF dan *cosine similarity*. Pendekatan ini digunakan karena metode statistik

berbasis pembobotan kata belum sepenuhnya mampu menangkap konteks makna dan pola linguistik eksplisit pada teks deskriptif berbahasa Indonesia. Sistem berbasis aturan terbukti efektif dalam menangkap pola linguistik dan pengetahuan domain melalui aturan heuristik dan pencocokan pola secara eksplisit (Pulungan, 2024). Dalam penelitian ini, *rule-based system* tidak mengubah nilai *cosine similarity*, melainkan digunakan untuk menyesuaikan kategori kesesuaian yang diperoleh pada tahap sebelumnya.

Proses validasi berbasis aturan hanya diterapkan pada entri dengan tingkat kesesuaian rendah hingga sedang (cosine similarity ≤ 0,69), sedangkan entri dengan nilai kesesuaian tinggi langsung dipertahankan kategorinya (Aubaid et al., 2024). Penyesuaian kategori dilakukan secara bertahap, di mana entri pada kategori *Tidak Sesuai* masih dapat naik ke tingkat *Cukup Sesuai* apabila memenuhi minimal satu indikator positif, guna mengakomodasi kasus ketika nilai *cosine similarity* rendah namun masih terdapat relevansi. Pendekatan ini sejalan dengan konsep *hybrid similarity* yang menekankan bahwa pengukuran kesamaan teks tidak seharusnya hanya bergantung pada satu metrik statistik, karena berpotensi mengabaikan keterkaitan makna yang bersifat kontekstual dan linguistik (Prasad, 2023). Selanjutnya, kenaikan ke tingkat *Sangat Sesuai* dibatasi hanya jika minimal tiga indikator positif terpenuhi dan maksimal satu tingkat kenaikan diperbolehkan, guna memastikan bahwa kategori tertinggi benar-benar mencerminkan kesesuaian makna yang kuat serta mencegah *overcorrection* akibat kemiripan parsial atau kebetulan, sebagaimana ditekankan dalam kajian *interpretable text similarity* yang

menyoroti pentingnya penggunaan bukti linguistik dan semantik yang memadai untuk menekan risiko *false positive* (Opitz et al., 2025).

Pendekatan hibrida ini sejalan dengan penelitian terbaru yang mengombinasikan pembobotan kata (TF-IDF), pengukuran kesamaan teks (*cosine similarity*), dan sistem berbasis aturan untuk memperkuat stabilitas hasil klasifikasi teks (Aubaid et al., 2024). Hasil dari TF-IDF dan *cosine similarity* kemudian dikoreksi dengan aturan linguistik untuk menghasilkan kategori akhir: Sesuai, Cukup Sesuai, atau Tidak Sesuai.

Evaluasi dilakukan dengan menganalisis distribusi skor kesesuaian, meninjau sampel manual beberapa entri untuk memverifikasi relevansi konteks budaya, serta memastikan setiap langkah analisis sesuai dengan kerangka CRISP-DM seperti dijelaskan oleh Sari et al. (2021) dalam *Jurnal Teknologi Informasi dan Komputer*.

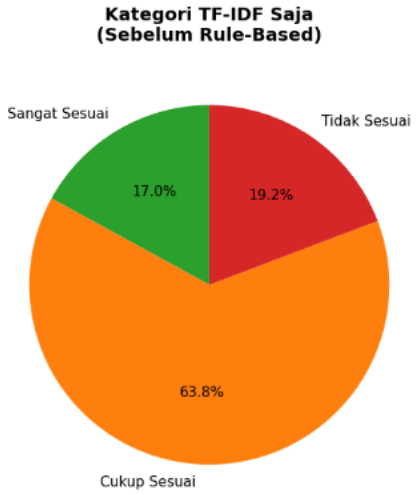
Dengan rancangan ini, proses pemodelan berjalan sistematis, transparan, dan dapat direplikasi oleh penelitian sejenis yang berfokus pada evaluasi kesesuaian teks budaya digital di Indonesia.

3.3.4 Evaluation & Visualization

Tahap ini bertujuan untuk mengevaluasi hasil analisis kesesuaian antara kolom *judul* dan *deskripsi* pada data Perpustakaan Digital Budaya Indonesia (PDBI) menggunakan kombinasi metode TF-IDF dan *rule-based validation*. Evaluasi dilakukan menggunakan data dari web budaya-indonesia.org sebanyak 58.709 kemudian di uji coba terhadap 1000 entri data yang diambil secara acak untuk mewakili karakteristik keseluruhan dataset. Tahapan ini berfokus pada

analisis nilai kesamaan teks *(similarity score)*, perubahan kategori hasil sebelum dan sesudah penerapan aturan tambahan (*rule-based*), serta visualisasi pola distribusi hasil.

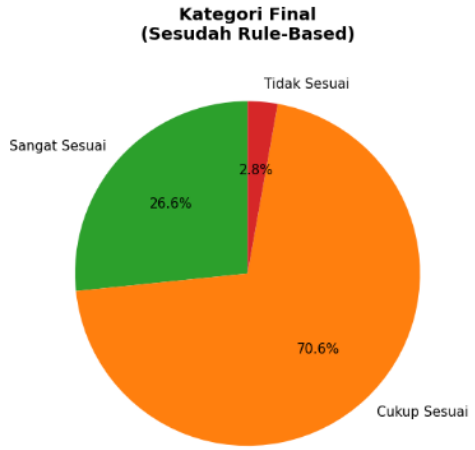
Langkah pertama adalah menghitung nilai *cosine similarity* berdasarkan pembobotan TF-IDF untuk setiap pasangan teks judul dan deskripsi. Nilai ini merepresentasikan tingkat kemiripan makna antar dua teks secara statistik tanpa mempertimbangkan konteks budaya.



Gambar 3. 3 Kategori TF-IDF

Hasil analisis awal menunjukkan bahwa sebagian besar entri data memiliki tingkat kesesuaian sedang, sebagaimana terlihat pada Gambar 3.3. Berdasarkan grafik tersebut, kategori *Cukup Sesuai* mendominasi dengan proporsi sebesar 63,8%, diikuti oleh *Sangat Sesuai* sebesar 17,0%, dan *Tidak Sesuai* sebesar 19,2%. Hal ini menunjukkan bahwa secara umum, deskripsi yang diunggah pengguna memiliki keterkaitan makna dengan judul, namun belum sepenuhnya menggambarkan relevansi semantik yang kuat dalam konteks kebudayaan.

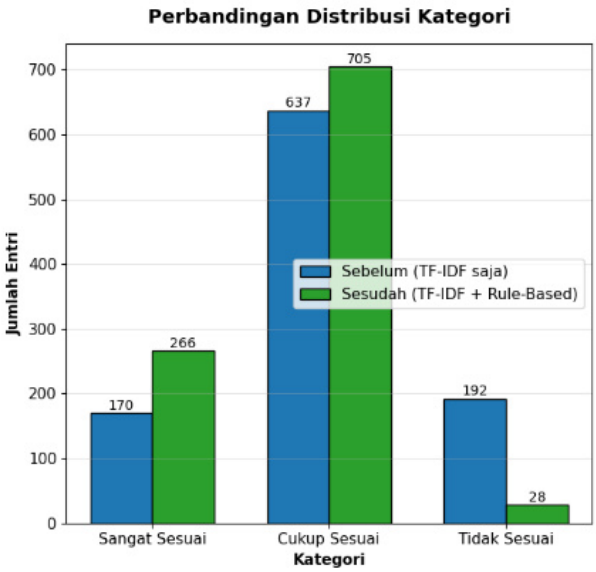
Pada tahap berikutnya, dilakukan penerapan sistem *rule-based validation* untuk memperkuat hasil TF-IDF dengan menambahkan lapisan penilaian berbasis aturan. Beberapa aturan yang digunakan meliputi pendeteksian konteks budaya, keberadaan sinonim dan istilah daerah, frasa deskriptif seperti "asal usul" atau "berfungsi sebagai", serta penalti untuk deskripsi yang terlalu pendek atau bersifat spam. Setelah penerapan *rule-based system*, hasil distribusi kategori berubah secara signifikan sebagaimana ditunjukkan pada Gambar 3.4.



Gambar 3. 4 Kategori Final

Kategori *Sangat Sesuai* meningkat menjadi 26,6%, *Cukup Sesuai* mencapai 70,6%, dan *Tidak Sesuai* menurun menjadi hanya 2,8%. Pergeseran proporsi ini menandakan bahwa pendekatan hibrida antara TF-IDF dan *rule-based* mampu meningkatkan ketepatan klasifikasi indikasi kesesuaian semantik, terutama dengan memberi kenaikan kategori pada deskripsi yang relevan secara kontekstual.

Perbandingan antara hasil TF-IDF murni dan hasil gabungan *rule-based* divisualisasikan dalam Gambar 3.5.



Gambar 3. 5 Perbandingan sebelum dan sesudah *rule based*

Grafik tersebut memperlihatkan penurunan jumlah entri pada kategori *Tidak Sesuai* dan peningkatan pada kategori *Cukup Sesuai* dan *Sangat Sesuai*. Hal ini menunjukkan bahwa *rule-based adjustment* berfungsi efektif dalam memperbaiki klasifikasi entri yang sebelumnya ambigu menjadi lebih sesuai dengan konteks budaya dan makna deskriptif yang terkandung.

Tabel 3. 14 Distribusi Kategori Sebelum dan Sesudah *Rule-based*

Kategori	***Baseline***	**Persentase**	**Final *(Hybrid)***	**Persentase**	**Perubahan Absolut**	**Perubahan (%)**
Sangat Sesuai	170	17,0%	266	26,6%	+96	+9,6%
Cukup Sesuai	637	63,8%	705	70,6%	+68	+6,8%
Tidak Sesuai	192	19,2%	28	2,8%	−164	−16,4%
Total	**999**	**100%**	**999**	**100%**	—	—

Pada tabel 3.14 perubahan menunjukkan kenaikan sebesar 96 entri pada kategori Sangat Sesuai dan 68 entri pada kategori Cukup Sesuai, serta penurunan sebesar 164 entri pada kategori Tidak Sesuai. Pergeseran ini mengindikasikan

bahwa sejumlah entri yang sebelumnya dinilai rendah secara statistik oleh TF-IDF ternyata memiliki indikator kontekstual yang relevan dan berhasil dikenali oleh sistem *rule-based*. Dengan demikian, distribusi akhir mencerminkan klasifikasi yang lebih proporsional dan lebih selaras dengan konteks semantik teks budaya.

Tabel 3. 15 Matriks Perubahan Kategori *Baseline* ke Final

***Baseline* ↓ / Final →**	**Sangat Sesuai**	**Cukup Sesuai**	**Tidak Sesuai**	**Total *Baseline***
Sangat Sesuai	170	0	0	170
Cukup Sesuai	96	541	0	637
Tidak Sesuai	0	164	28	192
Total Final	**266**	**705**	**28**	**999**

Tabel 3.15 menampilkan matriks perubahan kategori dari hasil *baseline* menuju hasil akhir setelah penerapan *rule-based validation.* Dari total 999 entri, sebanyak 739 entri (74%) tidak mengalami perubahan kategori, yang menunjukkan bahwa sistem *rule-based* tidak mengganggu struktur klasifikasi yang telah kuat pada *baseline* TF-IDF. Perubahan terjadi pada 260 entri (26%), yang sebagian besar merupakan peningkatan kategori.

Berdasarkan hasil evaluasi dan visualisasi tersebut, dapat disimpulkan bahwa integrasi metode TF-IDF dengan sistem *rule-based* memberikan peningkatan signifikan dalam pemisahan data valid dan tidak valid secara semantik. Entri yang semula dikategorikan *Cukup Sesuai* dapat naik menjadi *Sangat Sesuai* ketika mengandung sinonim atau frasa deskriptif yang menjelaskan konteks budaya tertentu, seperti “alat musik tradisional”, “pakaian adat”, atau “makanan khas daerah”. Sebaliknya, deskripsi yang terlalu pendek atau mengandung pengulangan kata tetap terklasifikasi sebagai *Tidak Sesuai*, sehingga hasil akhir tetap menjaga

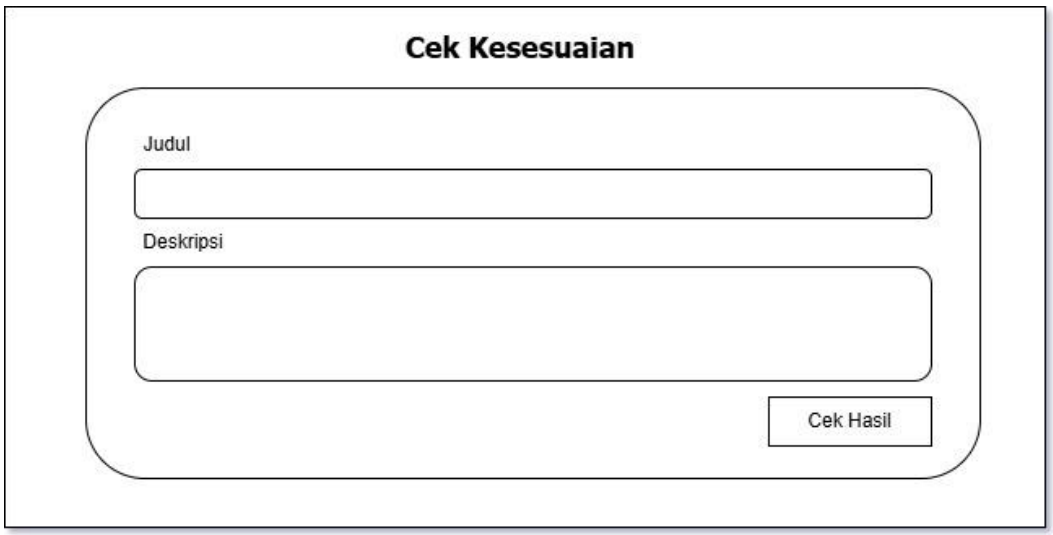
validitas dan konsistensi data. Secara keseluruhan, pendekatan *hybrid TF-IDF + rule-based* terbukti lebih efektif dalam menghasilkan analisis indikasi kesesuaian semantik yang tidak hanya mengandalkan kemiripan statistik, tetapi juga memperhatikan relevansi kontekstual dalam data budaya digital PDBI.

3.3.5 Implementasi Web

Sebagai tahap akhir dari penelitian ini, dikembangkan sebuah aplikasi web interaktif yang digunakan untuk melakukan analisis kesesuaian antara judul dan deskripsi data budaya secara otomatis. Aplikasi ini merupakan implementasi dari metode TF-IDF, cosine similarity, dan *rule-based* validation yang telah dibangun pada tahap sebelumnya. Sistem dikembangkan menggunakan bahasa pemrograman Python dengan *framework Streamlit*, sehingga aplikasi dapat dijalankan melalui antarmuka web secara interaktif dan mudah digunakan. Pada aplikasi ini, pengguna dapat memasukkan judul dan deskripsi budaya untuk dianalisis oleh sistem. Selanjutnya, sistem akan melakukan *preprocessing* teks, pembentukan vektor TF-IDF, perhitungan *cosine similarity*, dan validasi menggunakan *rule-based validation*. Hasil analisis kemudian ditampilkan dalam bentuk *similarity score* dan kategori kesesuaian teks.

Selain analisis satu data, aplikasi juga menyediakan fitur *batch processing* menggunakan file Excel sehingga pengguna dapat melakukan analisis terhadap banyak data secara otomatis. Hasil analisis dapat ditampilkan dalam bentuk tabel, visualisasi grafik, serta diekspor ke dalam format CSV dan Excel.

Gambar 3.6 berikut memperlihatkan tampilan desain *user interface* web hasil implementasi penelitian:



Gambar 3. 6 Halaman Input Judul dan Deskripsi

Desain UI pada Gambar 3.6 diatas memperkuat aspek praktis dan replikasi penelitian, karena memungkinkan model analisis digunakan secara langsung untuk memverifikasi kesesuaian teks budaya. Dengan demikian, aplikasi ini menjadi bukti nyata bahwa metode TF-IDF dan *rule-based system* dapat diimplementasikan secara efektif untuk mendukung validasi kualitas data budaya digital Indonesia.

3.4 Kriteria Evaluasi

Kriteria evaluasi digunakan untuk menilai sejauh mana sistem mampu mengukur kesesuaian semantik antara *judul* dan *deskripsi* secara akurat dan kontekstual. Evaluasi dilakukan melalui kombinasi analisis kuantitatif menggunakan nilai *cosine similarity* hasil perhitungan TF-IDF serta penyesuaian kualitatif melalui penerapan *rule-based validation*. Tahapan ini bertujuan untuk memastikan bahwa hasil analisis tidak hanya mengandalkan kemiripan tekstual,

tetapi juga mempertimbangkan makna budaya, sinonim, dan konteks deskriptif dari entri data budaya.

Pada tahap awal, model TF-IDF murni menghasilkan nilai *cosine similarity* dalam rentang 0 hingga 1. Setelah dilakukan observasi terhadap sebaran nilai, terlihat bahwa sebagian besar skor berada pada rentang menengah (0,20–0,69). Hal ini menunjukkan bahwa meskipun terdapat kemiripan kata atau istilah, kedalaman makna dan relevansi kontekstual belum sepenuhnya tertangkap oleh pendekatan berbasis statistik semata. Selanjutnya, dilakukan integrasi dengan *rule-based system* yang berfungsi sebagai mekanisme evaluasi tambahan. Sistem ini meningkatkan kategori pada deskripsi yang mengandung kata kunci dengan relevansi kontekstual tinggi (misalnya: "tarian tradisional", "makanan khas daerah", atau "upacara adat") serta memberikan penalti terhadap deskripsi yang terlalu singkat atau terindikasi spam.

Melalui evaluasi ulang setelah penerapan aturan tersebut, terjadi pergeseran distribusi skor, di mana sejumlah entri yang sebelumnya berada pada kategori Cukup Sesuai meningkat ke kategori Sangat Sesuai (≥ 0,70). Hal ini menunjukkan bahwa pendekatan kombinasi *(hybrid)* mampu meningkatkan sensitivitas model terhadap konteks dan kualitas deskripsi.

Tabel 3. 16 Kriteria Penilaian

Rentang Nilai *Cosine similarity*	Kategori Kesesuaian
≥ 0,70	Sangat Sesuai
0,20 – 0,69	Cukup Sesuai
< 0,20	Tidak Sesuai

Kategori pada Tabel 3.16 tidak ditentukan secara arbitrer, melainkan merupakan hasil dari proses evaluasi performa model dalam mengukur tingkat

kesesuaian semantik antar teks. Penetapan rentang nilai *cosine similarity* sebagai batas klasifikasi dilakukan berdasarkan analisis distribusi skor yang diperoleh dari pengujian terhadap 1000 data uji.

Melalui analisis tersebut, diamati pola sebaran nilai serta kecenderungan pengelompokan skor yang merepresentasikan tingkat kemiripan rendah, sedang, dan tinggi. Berdasarkan hasil observasi distribusi dan validasi terhadap sampel data, ditetapkan tiga kategori kesesuaian, yaitu sangat sesuai, cukup sesuai dan tidak sesuai. Berdasarkan hasil uji terhadap 1000 data, nilai *cosine similarity* dari model TF-IDF murni cenderung menghasilkan kategori *Cukup Sesuai* sebagai dominan, sedangkan setelah integrasi dengan *rule-based system*, banyak entri berpindah ke kategori *Sangat Sesuai*.

Tabel 3. 17 Contoh Perbandingan Data Berdasarkan Kategori Kesesuaian

No	**Judul**	**Deskripsi**	**Tf-idf**	**Kategori**	***Rule* Detail**
29	Arti dari Tari Dana-Dana di Gorontalo	Selain Tari Saronde, Tari Dana-danamerupakan salah satu dari seni budaya asli Gorontalo. Tari ini menampilkan gerakan yang harus diikuti oleh seluruh anggota badan dan menggambarkan pergaulan keakraban remaja. Salah satu tarian khas gorontalo yang biasanya ditarikan pada saat hajatan berupa acara perkawinan atau pesta rakyat dan pagelaran seni budaya. Keunikannya tari ini didominasi oleh gerakan-gerakan yang dinamis mengikuti irama musik gambus dan rebana serta lagu berisi pantun bertemakan percintaan, Dst....	0.4697	Cukup Sesuai	-
29	Arti dari Tari Dana-Dana di Gorontalo	Selain Tari Saronde, Tari Dana-danamerupakan salah satu dari seni budaya asli Gorontalo. Tari ini menampilkan gerakan yang	0.4697	Berubah menjadi Sangat Sesuai	4 *rule* positif: Overlap kata judul, Konteks budaya, Sinonim

No	Judul	Deskripsi	Tf-idf	Kategori	*Rule* Detail
		harus diikuti oleh seluruh anggota badan dan menggambarkan pergaulan keakraban remaja. Salah satu tarian khas gorontalo yang biasanya ditarikan pada saat hajatan berupa acara perkawinan atau pesta rakyat dan pagelaran seni budaya. Keunikannya tari ini didominasi oleh gerakan-gerakan yang dinamis mengikuti irama musik gambus dan rebana serta lagu berisi pantun bertemakan percintaan, Dst....			terdeteksi, Frasa deskriptif; Kategori Cukup Sesuai -> Sangat Sesuai; Alasan: TF-IDF 0.20-0.69 (0.470) + 4 *rule*s positif
42	Baju Adat Daerah Sumatera Utara – Pakaian Batak Pakpak	Baju Adat Daerah Sumatera Utara – Pakaian Batak Pakpak Etnis Batak Pakpak merujuk pada suku Batak Pakpak yang tinggal di daerah Kabupaten Pakpak Bharat dan Kabupaten Dairi. Baju adat Sumatera Utara dari daerah Pakpak ini menggunakan kain Oles, yaitu kain tradisional daerah Pakpak dilengkapi dengan berbagai aksesorisnya.	0,7641	Sangat Sesuai	-
42	Baju Adat Daerah Sumatera Utara – Pakaian Batak Pakpak	Baju Adat Daerah Sumatera Utara – Pakaian Batak Pakpak Etnis Batak Pakpak merujuk pada suku Batak Pakpak yang tinggal di daerah Kabupaten Pakpak Bharat dan Kabupaten Dairi. Baju adat Sumatera Utara dari daerah Pakpak ini menggunakan kain Oles, yaitu kain tradisional daerah Pakpak dilengkapi dengan berbagai aksesorisnya.	0,7641	Sangat Sesuai	TF-IDF > 0.69 (tidak perlu *rule-based*)
39	Baju Adat Daerah Sumatera Utara – Pakaian Batak Simalungun	Baju Adat Daerah Sumatera Utara – Pakaian Batak Simalungun Suku Batak memang miliki beberapa sub yang pada umumnya disesuaikan dengan posisi tempat tinggal mereka. Demikian pula dengan Suku Batak Simalungun. Suku Batak Simalungun memang kebanyakan tinggal di Kabupaten Simalungun dan	0.5568	Cukup Sesuai	-

No	Judul	Deskripsi	Tf-idf	Kategori	*Rule* Detail
		sekitarnya. Penggunaan kain ulos merupakan salah satu ciri pakaian adat Sumatera Utara. Dst….			
39	Baju Adat Daerah Sumatera Utara – Pakaian Batak Simalungun	Baju Adat Daerah Sumatera Utara – Pakaian Batak Simalungun Suku Batak memang miliki beberapa sub yang pada umumnya disesuaikan dengan posisi tempat tinggal mereka. Demikian pula dengan Suku Batak Simalungun. Suku Batak Simalungun memang kebanyakan tinggal di Kabupaten Simalungun dan sekitarnya. Penggunaan kain ulos merupakan salah satu ciri pakaian adat Sumatera Utara. Dst….	0.5568	Cukup Sesuai	1 *rule* positif: Overlap kata judul; TF-IDF 0.557 (0.20-0.69) -> butuh >=3 *rule*s (hanya 1)
7	Kotagede	Kota ini merupakan kawasan bersejarah yang merupakan The Old Capital City yang menyimpan sejarah mengenai lahirnya Mataram Islam . Berawal dari berdirinya sebuah kerajaan di tengah hutan pada tahun 1575 yang diprakarsai oleh Ki Ageng Pemanahan yang merupakan asal mula berdirinya kerajaan Mataram.	0.1891	Tidak Sesuai	-
7	Kotagede	Kota ini merupakan kawasan bersejarah yang merupakan The Old Capital City yang menyimpan sejarah mengenai lahirnya Mataram Islam . Berawal dari berdirinya sebuah kerajaan di tengah hutan pada tahun 1575 yang diprakarsai oleh Ki Ageng Pemanahan yang merupakan asal mula berdirinya kerajaan Mataram.	0.1891	Berubah Menjadi Cukup Sesuai	2 *rule* positif: Overlap kata judul, Konteks budaya; Kategori Tidak Sesuai -> Cukup Sesuai; Alasan: TF-IDF <= 0.20 (0.189) + 2 *rule* positif
319	Bebotok	Jawa tengah terkenal akan budayanya	0.0	Tidak Sesuai	-
319	Bebotok	Jawa tengah terkenal akan budayanya	0.0	Tidak Sesuai	Flag: Deskripsi pendek (<5 kata); Tidak ada *rule* positif terpenuhi
322	Test2	Begono (sumber: E-b ook Mahakarya 5000 Resep	0.0	Tidak Sesuai	-

No	Judul	Deskripsi	Tf-idf	Kategori	*Rule* Detail
		Makanan dan Minuman di Indonesia)			
322	Test2	Begono (sumber: E-b ook Mahakarya 5000 Resep Makanan dan Minuman di Indonesia)	0.0	Tidak Sesuai	Tidak ada *rule* yang berpengaruh

Data pada Tabel 3.17 menunjukkan perbedaan karakteristik antar kategori. Entri dengan kategori *Sangat Sesuai* memiliki indikasi keterkaitan semantik kuat antara judul dan deskripsi serta mengandung konteks budaya lokal yang jelas. Kategori *Cukup Sesuai* menunjukkan adanya hubungan sebagian, misalnya kata kunci yang sama tetapi penjelasan deskripsi tidak sepenuhnya mendukung konteks utama judul. Sementara itu, kategori *Tidak Sesuai* ditandai oleh deskripsi yang terlalu pendek, tidak relevan, atau tidak mengandung informasi semantik yang cukup untuk menjelaskan makna dari judul.

Hal ini menunjukkan bahwa evaluasi yang mempertimbangkan konteks linguistik dan semantik memberikan hasil yang lebih akurat dalam menilai kesesuaian antara judul dan deskripsi. Dengan demikian, kriteria evaluasi ini tidak hanya menilai kesamaan teks secara matematis, tetapi juga mengukur tingkat *semantic validity* dari setiap entri budaya, yang menjadi dasar bagi verifikasi kualitas data digital pada PDBI.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil

Pada bagian ini disajikan hasil dari seluruh tahapan penelitian yang telah dilakukan, mulai dari pengolahan data hingga penerapan metode yang digunakan. Hasil yang ditampilkan meliputi karakteristik data penelitian, proses pembentukan representasi teks menggunakan TF-IDF, perhitungan tingkat kesesuaian menggunakan *cosine similarity*, hingga implementasi *rule-based validation.* Penyajian hasil ini bertujuan untuk memberikan gambaran menyeluruh mengenai kinerja sistem dalam menganalisis kesesuaian antara judul dan deskripsi pada data teks budaya.

4.1.1 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data teks budaya yang diperoleh dari situs budaya-indonesia.org, yang merupakan bagian dari Perpustakaan Digital Budaya Indonesia (PDBI). *Platform* ini bersifat terbuka dan memungkinkan masyarakat untuk berkontribusi secara langsung dalam menambahkan informasi terkait artefak budaya. Oleh karena itu, data yang terkumpul merupakan hasil dari proses *crowdsourcing* yang memiliki tingkat keragaman tinggi, baik dari segi panjang teks, struktur bahasa, maupun kualitas informasi.

Jumlah data yang digunakan dalam penelitian ini adalah sebanyak 58.604 entri sejak November 2025 diambil dari situs budaya-indonesia.org sebagai sumber

utama informasi budaya digital, di mana setiap entri terdiri dari dua atribut utama yang digunakan, yaitu judul dan deskripsi. Judul berfungsi sebagai representasi utama dari nama artefak budaya, sedangkan deskripsi berisi uraian yang menjelaskan konteks, fungsi, atau makna budaya dari artefak tersebut. Kedua atribut ini menjadi fokus utama dalam penelitian karena digunakan untuk mengukur tingkat kesesuaian menggunakan metode TF-IDF dan *rule-based validation*.

Berdasarkan hasil eksplorasi data pada tahap *data understanding*, diperoleh gambaran mengenai karakteristik panjang teks pada masing-masing atribut. Berikut adalah contoh column yang ada di dataset.

Tabel 4. 1 Tampilan Dataset Budaya-Indonesia.org

No	Judul	Kategori	Elemen Budaya	Provinsi	Asal Daerah	Keterangan
1	Sejuta makna dari lagu Yamko Rambe Yamko	Lagu Adat	Musik dan Lagu	Papua	Irian Jaya	Makna Lagu Yamko Rambe Yamko - merupakan lagu yang berasal dari Irian Jaya/Papua..
2	Rumah Inggit Garnasih	Rumah	Produk Arsitektur	Jawa Barat	Bandung	Sumber : (http://cdn.metrotvnews...) Rumah bersejarah tersebut terlihat sangat bersih dan terawat, sebagaimana tempat tinggal di perumahan..
3	Nasihat Cari Jodoh dalam Lagu Maluku Huhate	Lagu Adat	Musik dan Lagu	Maluku	Ambon	Lagu daerah dari Maluku satu ini memiliki makna cukup dalam. Lagu ini berjudul "Huhate"..
4	Galeri Nasional Indonesia	Gedung	Produk Arsitektur	DKI Jakarta	Jakarta Pusat	Galeri Nasional Indonesia (bahasa Inggris: National Gallery of Indonesia) adalah sebuah lembaga budaya negara yang gedungnya antara

No	Judul	Kategori	Elemen Budaya	Provinsi	Asal Daerah	Keterangan
						lain berfungsi sebagai tempat pameran..
5	Komplek s Makam Imogiri	Makam	Produk Arsitektur	Daerah Istimewa Yogyakarta	Bantul	Permakaman Imogiri, Pasarean Imogiri, atau Pajimatan Girirejo Imogiri merupakan kompleks permakaman yang berlokasi di Imogiri, Bantul, DI Yogyakarta..

Berdasarkan Tabel 4.1, dataset yang digunakan dalam penelitian memiliki beberapa kolom utama, yaitu Judul, Kategori, Elemen Budaya, Provinsi, Asal Daerah, dan Keterangan. Kolom Judul digunakan sebagai representasi utama objek budaya, sedangkan kolom Keterangan berisi deskripsi atau penjelasan terkait objek tersebut. Sementara itu, kolom Kategori dan Elemen Budaya digunakan untuk menunjukkan jenis budaya, serta kolom Provinsi dan Asal Daerah menunjukkan lokasi asal budaya. Dalam penelitian ini, kolom yang digunakan sebagai fokus analisis adalah Judul dan Keterangan karena kedua kolom tersebut digunakan dalam proses pengukuran kesesuaian teks menggunakan metode TF-IDF dan *cosine similarity.*

Rangkuman statistik panjang teks pada dataset budaya-indonesia.org dapat dilihat pada Tabel 4.1 berikut.

Tabel 4. 2 Statistik Panjang Teks Data Penelitian

No	Atribut	Rata-rata	Minimum	Maksimum
1	Judul	23,9 karakter	1 karakter	229 karakter
2	Deskripsi	2056,5 karakter	1 karakter	32.765 karakter

Berdasarkan Tabel 4.2, terlihat bahwa terdapat perbedaan yang cukup signifikan antara panjang teks judul dan deskripsi. Judul cenderung memiliki

panjang yang relatif pendek dan ringkas, sedangkan deskripsi memiliki variasi yang sangat besar, mulai dari sangat singkat hingga sangat panjang. Ketimpangan ini menunjukkan bahwa kualitas dan kelengkapan informasi dalam data tidak seragam, yang berpotensi memengaruhi hasil analisis kesesuaian teks.

Selain variasi panjang teks, karakteristik data yang bersifat *crowdsourced* juga menimbulkan berbagai permasalahan kualitas data. Dalam proses eksplorasi, ditemukan bahwa sebagian entri memiliki deskripsi yang terlalu pendek sehingga tidak cukup representatif untuk menggambarkan konteks budaya. Di sisi lain, terdapat pula deskripsi yang sangat panjang namun mengandung pengulangan kata atau informasi yang tidak relevan. Penggunaan istilah daerah dan variasi bahasa juga menjadi karakteristik yang menonjol dalam dataset ini, yang menunjukkan kekayaan linguistik, namun sekaligus menjadi tantangan dalam proses analisis berbasis komputasi.

Untuk mengatasi berbagai permasalahan tersebut, dilakukan tahap *data preparation* sebelum proses pemodelan. Tahap ini bertujuan untuk membersihkan dan menormalkan teks agar memiliki struktur yang lebih seragam dan siap untuk dianalisis. Proses yang dilakukan meliputi *case folding* untuk menyeragamkan huruf menjadi huruf kecil, *tokenizing* untuk memecah teks menjadi unit kata, *stopword removal* untuk menghilangkan kata-kata umum yang tidak memiliki makna signifikan, serta *stemming* untuk mengubah kata menjadi bentuk dasarnya. Selain itu, dilakukan pula proses pembersihan terhadap karakter non-linguistik seperti angka, tanda baca, dan simbol yang tidak relevan.

Hasil dari proses *preprocessing* ini menghasilkan data yang lebih terstruktur dan konsisten, sehingga dapat direpresentasikan dalam bentuk numerik menggunakan metode TF-IDF pada tahap selanjutnya. Contoh hasil *preprocessing* ditampilkan pada Tabel 4.3 berikut.

Tabel 4. 3 Contoh Data Setelah *Preprocessing*

No	Judul	Judul_processed	Deskripsi	Deskripsi_processed
1	Sejuta makna dari lagu Yamko Rambe Yamko	juta makna lagu yamko rambe yamko	Makna Lagu Yamko Rambe Yamko - merupakan lagu yang berasal dari Irian Jaya/Papua.	makna lagu yamko rambe yamko rupa lagu asal iri jaya papua
2	Rumah Inggit Garnasih	rumah inggit garnasih	Sumber : (http://cdn.metrotvnews...) Rumah bersejarah tersebut terlihat sangat bersih dan terawat,	sumber rumah sejarah sebut lihat sangat bersih..
3	Nasihat Cari Jodoh dalam Lagu Maluku Huhate	nasihat cari jodoh lagu malu huhate	Lagu daerah dari Maluku satu ini memiliki makna cukup dalam. Lagu ini berjudul "Huhate".	lagu daerah maluku satu milik makna cukup lagu judul huhate
4	Galeri Nasional Indonesia	galeri nasional indonesia	Galeri Nasional Indonesia (bahasa Inggris: National Gallery of Indonesia) adalah sebuah lembaga budaya negara	galeri nasional indonesia bahasa inggris national gallery of indonesia buah lembaga budaya negara
5	Kompleks Makam Imogiri	kompleks makam imogiri	Permakaman Imogiri, Pasarean Imogiri, atau Pajimatan Girirejo Imogiri merupakan kompleks permakaman yang berlokasi di Imogiri, Bantul, DI Yogyakarta.	makam imogiri pasarean imogiri pajimatan girirejo imogiri rupa kompleks makam lokasi imogiri bantul yogyakarta

Melalui proses ini, data yang semula memiliki variasi tinggi dalam penulisan dan struktur bahasa menjadi lebih seragam dan siap untuk dianalisis. Hal ini sangat penting karena kualitas data yang baik akan berpengaruh langsung terhadap akurasi hasil perhitungan TF-IDF dan *cosine similarity*.

Secara keseluruhan, data penelitian ini mencerminkan kondisi nyata dari pengelolaan data budaya digital berbasis kontribusi publik yang memiliki keunggulan dalam jumlah dan keberagaman, namun juga menghadapi tantangan dalam hal konsistensi dan kualitas informasi. Oleh karena itu, tahap pemahaman dan persiapan data menjadi bagian krusial dalam penelitian ini untuk memastikan bahwa proses analisis dapat menghasilkan temuan yang valid dan representatif.

4.1.2 Hasil Perhitungan TF-IDF

Pada tahap ini dilakukan proses representasi data teks ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Tujuan dari proses ini adalah untuk mengubah data teks yang telah melalui tahap *preprocessing* menjadi vektor numerik, sehingga hubungan antar teks dapat dianalisis secara matematis. Representasi ini menjadi dasar dalam menghitung tingkat kesesuaian antara judul dan deskripsi pada data budaya.

Pembentukan vektor TF-IDF dilakukan dengan menggabungkan seluruh teks dari atribut judul dan deskripsi ke dalam satu korpus. Penggabungan ini bertujuan untuk memastikan bahwa kedua jenis teks berada dalam ruang fitur yang sama, sehingga dapat dibandingkan secara langsung menggunakan *cosine similarity*. Dalam proses ini digunakan *TF-IDF Vectorizer*, yaitu sebuah fungsi/library dari Scikit-learn (sklearn) yang digunakan untuk mengubah teks menjadi vektor numerik. Metode ini memberi bobot pada setiap kata berdasarkan tingkat kemunculannya dalam dokumen dibandingkan dengan dokumen lain.

Konfigurasi *ngram_range = (1,2)* digunakan agar sistem dapat mempertimbangkan unigram (kata tunggal) dan bigram (dua kata), sehingga

representasi teks menjadi lebih kaya dan mampu menangkap konteks frasa. Untuk menjaga efisiensi komputasi serta menghindari dominasi kata yang jarang muncul, jumlah fitur dibatasi sebanyak 10.000 kata yang paling relevan berdasarkan frekuensi kemunculannya dalam korpus. Pembatasan ini dilakukan agar model tetap mampu merepresentasikan informasi penting tanpa meningkatkan kompleksitas secara berlebihan. Hasil dari proses pembentukan vektor menghasilkan matriks TF-IDF dengan dimensi pada Tabel 4.4 sebagai berikut:

Tabel 4. 4 Dimensi Matriks TF-IDF

No	**Parameter**	**Nilai**
1	Jumlah Dokumen	117.208
2	Jumlah Fitur	10.000

Berdasarkan Tabel 4.4 jumlah dokumen sebesar 117.208 diperoleh dari penggabungan teks pada atribut judul dan deskripsi, di mana setiap entri data terdiri dari satu teks judul dan satu teks deskripsi yang masing-masing direpresentasikan sebagai dokumen dalam proses TF-IDF. Setiap dokumen kemudian direpresentasikan sebagai vektor berdimensi 10.000, di mana setiap elemen dalam vektor menunjukkan bobot TF-IDF dari suatu kata terhadap dokumen tersebut. Representasi dalam bentuk matriks ini memungkinkan proses perhitungan kesamaan antar teks dilakukan secara efisien pada tahap selanjutnya, khususnya dalam perhitungan *cosine similarity*. Setelah proses pembentukan vektor selesai, setiap kata dalam dokumen memiliki bobot TF-IDF yang mencerminkan tingkat kepentingannya dalam dokumen tersebut. Nilai TF-IDF diperoleh dari kombinasi dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Komponen TF menunjukkan frekuensi kemunculan suatu kata dalam sebuah dokumen. Semakin sering suatu kata muncul dalam dokumen, maka nilai TF akan semakin tinggi. Namun, frekuensi kemunculan saja tidak cukup untuk menentukan kepentingan suatu kata, karena beberapa kata umum dapat muncul di banyak dokumen. Oleh karena itu, digunakan komponen IDF yang berfungsi untuk mengurangi bobot kata-kata yang terlalu sering muncul di seluruh dokumen. Kata yang muncul di banyak dokumen akan memiliki nilai IDF yang rendah, sedangkan kata yang jarang muncul akan memiliki nilai IDF yang lebih tinggi. Dengan demikian, kombinasi TF dan IDF menghasilkan bobot yang lebih representatif terhadap kepentingan kata dalam konteks dokumen. Dalam konteks penelitian ini, kata-kata yang memiliki bobot TF-IDF tinggi umumnya merupakan kata kunci yang secara spesifik menggambarkan artefak budaya, seperti nama tarian, makanan, atau istilah khas daerah. Sebagai contoh, pada entri dengan judul “Sejuta makna dari lagu Yamko Rambe Yamko”, kata “yamko” memiliki bobot TF-IDF yang paling tinggi karena merupakan kata yang paling spesifik dan paling merepresentasikan isi dokumen. Selain itu, kombinasi kata seperti “lagu yamko”, “rambe yamko”, dan “yamko rambe” juga memiliki bobot yang cukup tinggi, yang menunjukkan bahwa sistem mampu menangkap konteks frasa melalui penggunaan unigram dan bigram. Sebaliknya, kata-kata yang bersifat lebih umum seperti “lagu” memiliki bobot yang lebih rendah karena muncul pada lebih banyak dokumen sehingga nilai IDF yang dihasilkan menjadi lebih kecil.

Berdasarkan hasil pembobotan TF-IDF, setiap dokumen memiliki kata-kata dengan bobot yang berbeda sesuai tingkat kepentingannya terhadap isi dokumen.

Semakin spesifik suatu kata dan semakin jarang muncul pada dokumen lain, maka bobot TF-IDF yang dihasilkan akan semakin tinggi. Sebagai ilustrasi, contoh hasil pembobotan TF-IDF pada dokumen dengan judul “Sejuta makna dari lagu Yamko Rambe Yamko” ditampilkan pada Tabel 4.5 berikut.

Tabel 4. 5 Contoh Hasil Pembobotan TF-IDF

No	**Kata**	**Bobot TF-IDF**
1	yamko	0.520964
2	juta makna	0.326015
3	lagu yamko	0.307690
4	rambe	0.307690
5	rambe yamko	0.307690
6	yamko rambe	0.307690
7	juta	0.294688
8	makna lagu	0.276364
9	makna	0.234952
10	lagu	0.158737

Berdasarkan Tabel 4.5, terlihat bahwa kata “yamko” memiliki bobot TF-IDF tertinggi dibandingkan kata lainnya. Hal ini menunjukkan bahwa kata tersebut merupakan kata yang paling representatif dalam dokumen dan memiliki tingkat kekhususan yang tinggi dibandingkan dokumen lain pada dataset. Selain itu, munculnya kombinasi kata seperti “lagu yamko”, “rambe yamko”, dan “yamko rambe” menunjukkan bahwa penggunaan konfigurasi ngram_range = (1,2) memungkinkan sistem untuk menangkap unigram maupun bigram, sehingga representasi konteks frasa menjadi lebih baik. Kata-kata umum seperti “lagu” memiliki bobot yang lebih rendah dibandingkan istilah yang lebih spesifik. Hal ini menunjukkan bahwa metode TF-IDF mampu memberikan penekanan lebih besar pada kata atau frasa yang memiliki kontribusi penting dalam isi dokumen.

Hasil pembobotan ini memberikan representasi numerik yang lebih bermakna dibandingkan sekadar kemunculan kata, karena mempertimbangkan

tingkat kepentingan kata dalam keseluruhan korpus. Dengan demikian, model dapat lebih fokus pada kata-kata yang benar-benar relevan dalam menggambarkan isi dokumen. Meskipun demikian, metode TF-IDF memiliki keterbatasan dalam menangkap hubungan semantik yang tidak bersifat eksplisit. TF-IDF tidak dapat mengenali sinonim, variasi bahasa daerah, maupun konteks budaya yang tidak secara langsung disebutkan dalam teks. Akibatnya, dua teks yang memiliki makna serupa namun menggunakan kata yang berbeda dapat memiliki nilai similarity yang rendah. Oleh karena itu, hasil pembobotan TF-IDF dalam penelitian ini digunakan sebagai dasar awal (*baseline*) dalam analisis kesesuaian teks, yang selanjutnya akan disempurnakan melalui pendekatan *rule-based validation* untuk meningkatkan sensitivitas terhadap konteks semantik.

4.1.3 Hasil Pengukuran *Cosine Similarity*

Setelah dilakukan pembobotan teks menggunakan metode TF-IDF, tahap selanjutnya adalah menghitung tingkat kesesuaian antara judul dan deskripsi menggunakan metode *cosine similarity*. Metode ini digunakan untuk mengukur tingkat kemiripan antara dua vektor dalam ruang multidimensi, yaitu vektor judul dan vektor deskripsi pada setiap entri data. Perhitungan *cosine similarity* menghasilkan nilai dalam rentang 0 hingga 1. Nilai ini menunjukkan tingkat kedekatan antara dua teks, di mana nilai yang lebih tinggi menunjukkan tingkat kesesuaian yang lebih kuat.

Berdasarkan hasil perhitungan terhadap seluruh data, diperoleh bahwa nilai cosine similarity berada dalam rentang 0,000 hingga 1,000. Rentang ini menunjukkan bahwa dataset memiliki variasi tingkat kesesuaian yang sangat luas,

mulai dari tidak memiliki kesamaan sama sekali hingga memiliki kemiripan yang sangat tinggi. Nilai *similarity* mendekati 0 menunjukkan bahwa tidak terdapat kesamaan kata yang signifikan antara judul dan deskripsi. Kondisi ini umumnya terjadi pada data yang memiliki deskripsi kosong, terlalu singkat, atau tidak relevan dengan judul. Sebaliknya, nilai yang mendekati 1 menunjukkan bahwa terdapat kemiripan yang sangat tinggi antara kedua teks, biasanya karena adanya kesamaan kata atau frasa yang dominan.

Sebagai ilustrasi, beberapa contoh nilai *similarity* yang dihasilkan dalam penelitian ini ditampilkan pada Tabel 4.6 berikut.

Tabel 4. 6 Contoh Nilai *Cosine Similarity*

No	Judul_processed	Deskripsi_processed	Nilai Similarity
1	juta makna lagu yamko rambe yamko	makna lagu yamko rambe yamko rupa lagu asal iri jaya papua	0.318642
2	rumah inggit garnasih	sumber rumah sejarah sebut lihat sangat bersih..	0.202048
3	nasihat cari jodoh lagu malu huhate	lagu daerah maluku satu milik makna cukup lagu judul huhate	0.212456
4	galeri nasional indonesia	galeri nasional indonesia bahasa inggris national gallery of indonesia buah lembaga budaya negara	0.314624
5	kompleks makam imogiri	makam imogiri pasarean imogiri pajimatan girirejo imogiri rupa kompleks makam lokasi imogiri bantul yogyakarta	0.237953

Berdasarkan contoh pada Tabel 4.6, terlihat bahwa sebagian besar nilai similarity berada pada rentang menengah, yaitu sekitar 0,20 hingga 0,60 (Bashir et al., 2021). Hal ini menunjukkan bahwa sebagian besar pasangan judul dan deskripsi memiliki keterkaitan yang bersifat parsial, namun belum cukup kuat untuk dikategorikan sebagai sangat sesuai. Selain itu, keberadaan nilai 0 pada beberapa data menunjukkan bahwa metode TF-IDF tidak menemukan kesamaan kata antara

judul dan deskripsi setelah proses *preprocessing.* Hal ini mengindikasikan adanya permasalahan kualitas data, seperti deskripsi yang tidak relevan atau penggunaan istilah yang berbeda secara leksikal namun memiliki makna yang sebenarnya berkaitan.

Dengan demikian, distribusi nilai *similarity* yang cenderung berada pada rentang menengah menunjukkan bahwa pendekatan TF-IDF mampu menangkap kesamaan dasar antar teks, namun masih memiliki keterbatasan dalam memahami konteks semantik yang lebih dalam.

Untuk mempermudah interpretasi hasil, nilai *cosine similarity* yang diperoleh kemudian dikategorikan ke dalam tiga tingkat kesesuaian, yaitu *Tidak Sesuai*, *Cukup Sesuai*, dan *Sangat Sesuai*. Kategori ini ditentukan berdasarkan ambang batas (*threshold*) yang telah ditetapkan (Bashir et al., 2021). Rentang kategori kesesuaian dapat dilihat pada Tabel 4.7 berikut.

Tabel 4. 7 Kategori Kesesuaian Berdasarkan *Cosine Similarity*

Rentang Nilai	**Kategori**	**Interpretasi**
< 0,20	Tidak Sesuai	Tidak terdapat keterkaitan makna yang jelas antara judul dan deskripsi
0,20 – 0,69	Cukup Sesuai	Terdapat keterkaitan sebagian antara judul dan deskripsi
≥ 0,70	Sangat Sesuai	Memiliki kesesuaian makna yang kuat

Berdasarkan hasil klasifikasi terhadap seluruh data, diperoleh distribusi kategori sebagai berikut:

Tabel 4. 8 Distribusi Kategori Kesesuaian (*Baseline* TF-IDF)

Kategori	**Jumlah**	**Persentase**
Sangat Sesuai	3532	6%
Cukup Sesuai	31.234	53,3%
Tidak Sesuai	23.838	40,7%
Total	**58.604**	**100%**

Berdasarkan Tabel 4.8, terlihat bahwa mayoritas data berada pada kategori *Cukup Sesuai*, yaitu sebesar 53,3%. Sementara itu, jumlah data dalam kategori *Sangat Sesuai* sangat kecil, yaitu hanya 6%. Di sisi lain, hampir setengah dari data termasuk dalam kategori *Tidak Sesuai*. Distribusi ini menunjukkan bahwa metode TF-IDF cenderung menghasilkan nilai kesesuaian yang berada pada tingkat menengah. Hal ini disebabkan oleh karakteristik metode yang hanya mempertimbangkan kemunculan kata secara literal tanpa memahami hubungan makna yang lebih kompleks. Akibatnya, banyak pasangan teks yang sebenarnya relevan secara semantik, namun tidak terdeteksi sebagai sangat sesuai karena menggunakan istilah yang berbeda.

Dengan demikian, hasil pengukuran *cosine similarity* ini menjadi dasar awal dalam evaluasi kesesuaian teks. Namun, keterbatasan yang dimiliki oleh pendekatan ini menunjukkan perlunya metode tambahan yang mampu menangkap konteks semantik secara lebih mendalam, yang dalam penelitian ini dilakukan melalui penerapan *rule-based validation* pada tahap selanjutnya.

4.1.4 Hasil Implementasi *Rule-Based Validation*

Setelah diperoleh hasil *baseline* menggunakan TF-IDF dan *cosine similarity,* tahap selanjutnya adalah penerapan *rule-based validation* untuk memperbaiki hasil klasifikasi kesesuaian teks. Pendekatan ini dirancang untuk melengkapi keterbatasan TF-IDF yang hanya mempertimbangkan kemunculan kata secara literal, dengan menambahkan aturan berbasis konteks linguistik dan domain budaya.

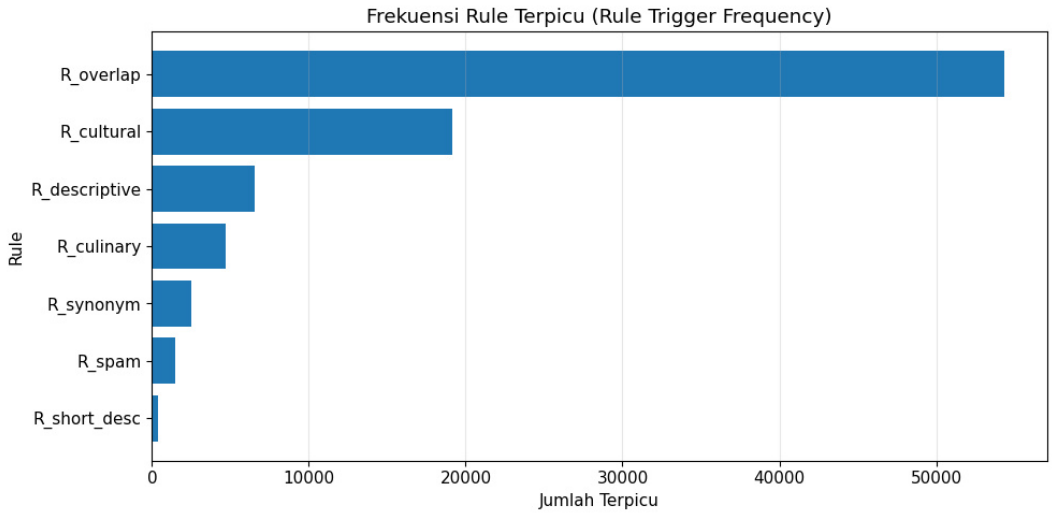
Sistem *rule-based* dalam penelitian ini terdiri dari beberapa aturan yang dirancang untuk mendeteksi indikasi kesesuaian antara judul dan deskripsi berdasarkan karakteristik teks. Aturan-aturan tersebut meliputi kesamaan kata, keberadaan konteks budaya, indikasi kuliner, frasa deskriptif, deteksi sinonim, serta aturan penalti seperti deskripsi yang terlalu pendek dan indikasi spam. Secara umum, aturan yang digunakan dalam penelitian ini dapat dirangkum pada Tabel 4.9 sebagai berikut:

Tabel 4. 9 Daftar Aturan *Rule-based*

No	**Nama *Rule***	**Deskripsi**
1	Overlap Kata	Mendeteksi kesamaan kata antara judul dan deskripsi
2	Konteks Budaya	Mendeteksi frasa yang berkaitan dengan budaya
3	Relevansi Kuliner	Mendeteksi konteks makanan berdasarkan indikator memasak
4	Sinonim	Mendeteksi kesamaan makna menggunakan sinonim
5	Frasa Deskriptif	Mendeteksi frasa yang menunjukkan penjelasan objek
6	Deskripsi Pendek	Memberikan penalti pada deskripsi yang terlalu singkat
7	Spam/Repetisi	Mendeteksi pengulangan kata yang tidak wajar

Penerapan aturan dilakukan secara selektif, di mana *rule-based* hanya diterapkan pada data dengan nilai TF-IDF di bawah 0,70. Hal ini bertujuan untuk menghindari perubahan pada data yang sudah memiliki tingkat kesesuaian tinggi berdasarkan TF-IDF. Selain itu, sistem dirancang dengan mekanisme *incremental adjustment*, di mana kategori kesesuaian hanya dapat naik satu tingkat berdasarkan jumlah aturan positif yang terpenuhi. Pendekatan ini digunakan untuk menjaga keseimbangan antara fleksibilitas model dan stabilitas hasil klasifikasi.

Setelah seluruh aturan diterapkan pada dataset, dilakukan analisis terhadap frekuensi *rule* yang terpicu. Hasil ini divisualisasikan dalam bentuk diagram batang pada Gambar 4.1.



Gambar 4. 1 Frekuensi *Rule Based* Terpicu

Berdasarkan hasil analisis, diperoleh jumlah trigger untuk masing-masing *rule* sebagai berikut:

Tabel 4. 10 Frekuensi *Rule* Terpicu

Rule	**Jumlah Trigger**
R_overlap	54.365
R_cultural	19.168
R_descriptive	6.527
R_culinary	4.707
R_synonym	2.488
R_spam	1.456
R_short_desc	355

Berdasarkan Tabel 4.10 dan visualisasi pada Gambar 4.1, terlihat bahwa *rule* overlap kata (R_overlap) merupakan *rule* yang paling dominan, dengan jumlah trigger mencapai 54.365 kali. Hal ini menunjukkan bahwa sebagian besar data memiliki kesamaan kata antara judul dan deskripsi, meskipun kesamaan tersebut belum tentu menghasilkan nilai similarity yang tinggi dalam TF-IDF.

Selanjutnya, *rule* konteks budaya (R_cultural) juga memiliki frekuensi yang cukup tinggi, yaitu sebesar 19.168. Hal ini menunjukkan bahwa banyak deskripsi

mengandung frasa yang berkaitan dengan konteks budaya, seperti “tradisional”, “upacara adat”, atau “warisan budaya”. Keberadaan frasa ini membantu sistem dalam mengidentifikasi kesesuaian yang tidak terdeteksi oleh pendekatan berbasis frekuensi kata.

Rule frasa deskriptif (R_descriptive) dan relevansi kuliner (R_culinary) juga memberikan kontribusi yang cukup signifikan dalam proses validasi. Kedua *rule* ini membantu sistem dalam mengenali pola deskripsi yang bersifat informatif, terutama pada data yang berkaitan dengan makanan atau artefak budaya tertentu.

Di sisi lain, *rule* sinonim (R_synonym) memiliki frekuensi yang lebih rendah, yang menunjukkan bahwa variasi kata dengan makna yang sama tidak terlalu dominan dalam dataset. Sementara itu, *rule* spam (R_spam) dan deskripsi pendek (R_short_desc) berfungsi sebagai mekanisme kontrol kualitas, dengan jumlah trigger yang relatif kecil.

Jika divisualisasikan, diagram batang pada Gambar 4.1 menunjukkan bahwa distribusi frekuensi *rule* bersifat tidak merata, di mana sebagian besar kontribusi berasal dari beberapa *rule* utama. Hal ini mengindikasikan bahwa tidak semua *rule* memiliki pengaruh yang sama dalam proses validasi, sehingga kombinasi *rule* menjadi faktor penting dalam menentukan hasil akhir.

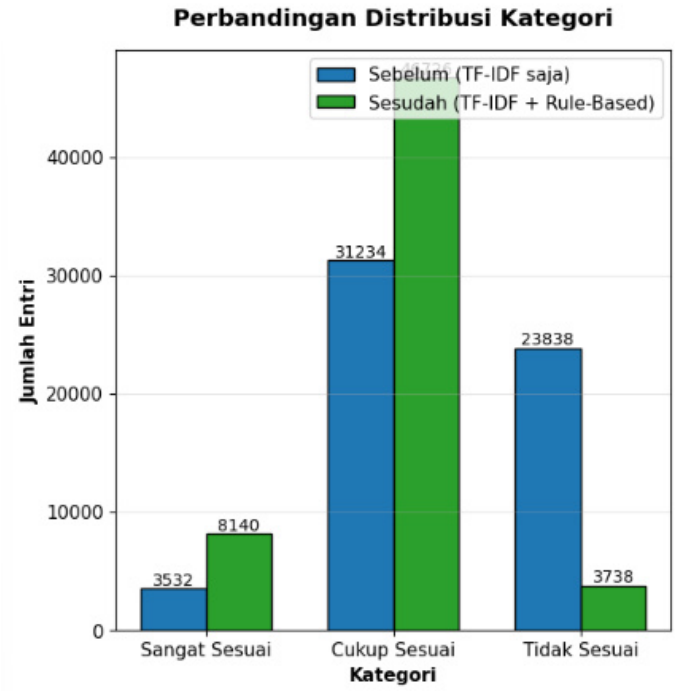
Secara keseluruhan, hasil penerapan *rule-based* menunjukkan bahwa pendekatan ini mampu menambahkan lapisan analisis yang tidak dimiliki oleh TF-IDF, terutama dalam menangkap konteks budaya dan pola deskriptif dalam teks. Dengan demikian, *rule-based* validation berperan sebagai mekanisme

penyempurnaan yang meningkatkan sensitivitas model dalam mengidentifikasi kesesuaian teks.

4.1.5 Perbandingan Hasil Sebelum dan Sesudah *Rule-Based*

Pada bagian ini dilakukan perbandingan antara hasil klasifikasi menggunakan metode TF-IDF (*baseline*) dengan hasil setelah penerapan *rule-based validation*. Perbandingan ini bertujuan untuk melihat sejauh mana *rule-based* mampu memperbaiki hasil klasifikasi dengan mempertimbangkan konteks linguistik dan budaya yang tidak terdeteksi oleh TF-IDF.

Perubahan distribusi kategori sebelum dan sesudah penerapan *rule-based* divisualisasikan dalam bentuk diagram batang pada Gambar 4.2.



Gambar 4. 2 Perubahan Distribusi Kategori Sebelum dan Sesudah Penerapan *Rule Based*

Berdasarkan hasil analisis, diperoleh perubahan distribusi kategori sebagai berikut:

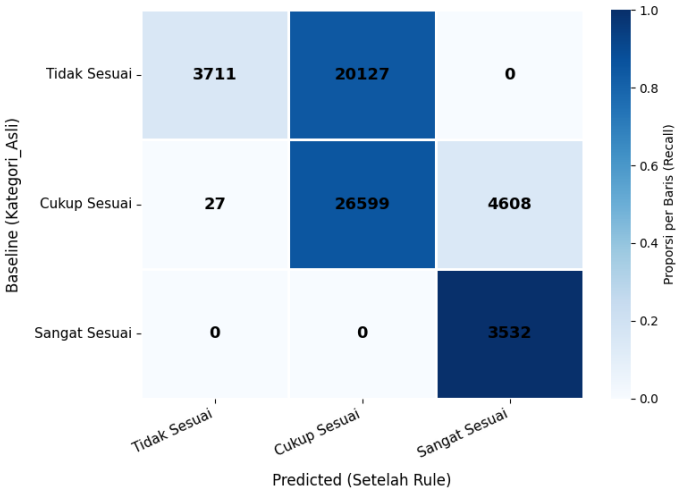
Tabel 4. 11 Perbandingan Distribusi Kategori

Kategori	**Sesudah (TF-IDF+*Cosine Similarity* +*Rule-based*)**	**Sesudah (TF-IDF+*Cosine Similarity* +*Rule-based*)**
Sangat Sesuai	3.532	8.140
Cukup Sesuai	31.234	46.726
Tidak Sesuai	23.838	3.738
Total	**58.604**	**58.604**

Berdasarkan Tabel 4.11 dan visualisasi pada Gambar 4.2, terlihat bahwa terjadi perubahan yang sangat signifikan pada distribusi kategori. Jumlah data dalam kategori *Tidak Sesuai* mengalami penurunan drastis dari 23.838 menjadi 3.738 entri. Sebaliknya, kategori *Cukup Sesuai* dan *Sangat Sesuai* mengalami peningkatan yang cukup besar.

Peningkatan paling mencolok terjadi pada kategori *Sangat Sesuai*, yang meningkat dari hanya 3.532 menjadi 8.140 entri. Hal ini menunjukkan bahwa *rule-based validation* mampu mengidentifikasi kesesuaian yang sebelumnya tidak terdeteksi oleh TF-IDF. Secara keseluruhan, perubahan ini menunjukkan bahwa pendekatan *rule-based* memberikan dampak signifikan dalam meningkatkan sensitivitas model terhadap kesesuaian teks.

Untuk menganalisis perubahan klasifikasi secara lebih rinci, digunakan matriks perubahan kategori dengan menjadikan hasil TF-IDF sebagai *baseline*. Visualisasi matriks perubahan kategori ditampilkan pada Gambar 4.3.



Gambar 4. 3 Matriks Perubahan Kategori *Baseline* TF-IDF vs *Rule Based*

Hasil matriks perubahan kategori ditunjukkan pada Tabel 4.12 berikut:

Tabel 4. 12 Matriks Perubahan Kategori (*Baseline* vs *Rule-based*)

TF-IDF ↓ / *Rule-based* →	**Tidak Sesuai**	**Cukup Sesuai**	**Sangat Sesuai**
True Tidak Sesuai	3.711	20.127	0
True Cukup Sesuai	27	26.599	4.608
True Sangat Sesuai	0	0	3532

Berdasarkan Tabel 4.12, terlihat bahwa sebagian besar data yang sebelumnya dikategorikan sebagai *Tidak Sesuai* oleh TF-IDF berpindah ke kategori *Cukup Sesuai*, yaitu sebanyak 20.127 entri. Hal ini menunjukkan bahwa *rule-based validation* memiliki kemampuan untuk meningkatkan tingkat kesesuaian dengan mempertimbangkan konteks tambahan. Selain itu, terdapat 4.608 data yang berpindah dari kategori *Cukup Sesuai* ke *Sangat Sesuai*, yang menunjukkan bahwa *rule-based* mampu memperkuat keyakinan model terhadap kesesuaian yang lebih tinggi.

Perubahan klasifikasi tersebut juga menunjukkan adanya perbedaan hasil antara *baseline* TF-IDF dan model setelah penerapan *rule-based*. Perbedaan ini

tidak menunjukkan penurunan kualitas model, melainkan menunjukkan bahwa *rule-based* melakukan penyesuaian klasifikasi dengan mempertimbangkan aspek kontekstual yang tidak sepenuhnya dapat ditangkap oleh pendekatan berbasis frekuensi kata.

Untuk memahami perubahan klasifikasi secara lebih mendalam, dilakukan analisis terhadap pola perpindahan kategori setelah penerapan *rule-based*. Perubahan yang paling dominan adalah perpindahan dari kategori Tidak Sesuai ke Cukup Sesuai. Kondisi ini menunjukkan bahwa banyak data yang sebelumnya dianggap kurang relevan oleh TF-IDF sebenarnya memiliki hubungan makna yang tidak terdeteksi karena keterbatasan metode berbasis kesamaan kata. Sebagai contoh, suatu deskripsi dapat menjelaskan objek budaya menggunakan istilah yang berbeda dari judul sehingga tidak memiliki kata yang sama secara langsung. Dalam kondisi tersebut, *rule-based* mampu mendeteksi keterkaitan melalui indikator tambahan seperti konteks budaya, pola frasa deskriptif, maupun istilah yang berkaitan dengan objek budaya tertentu.

Selain itu, terdapat pula perubahan dari kategori Cukup Sesuai menjadi Sangat Sesuai. Perubahan ini umumnya terjadi pada data yang memenuhi beberapa indikator secara bersamaan, seperti kesamaan istilah, keterkaitan konteks budaya, dan keberadaan frasa deskriptif tertentu. Kombinasi indikator tersebut memperkuat tingkat kesesuaian antara judul dan deskripsi sehingga kategori klasifikasi meningkat menjadi Sangat Sesuai. Di sisi lain, perubahan ke arah penurunan kategori relatif jarang terjadi, yang menunjukkan bahwa *rule-based* lebih banyak

berperan dalam meningkatkan sensitivitas sistem terhadap keterkaitan makna dibandingkan melakukan pengurangan tingkat kesesuaian secara agresif.

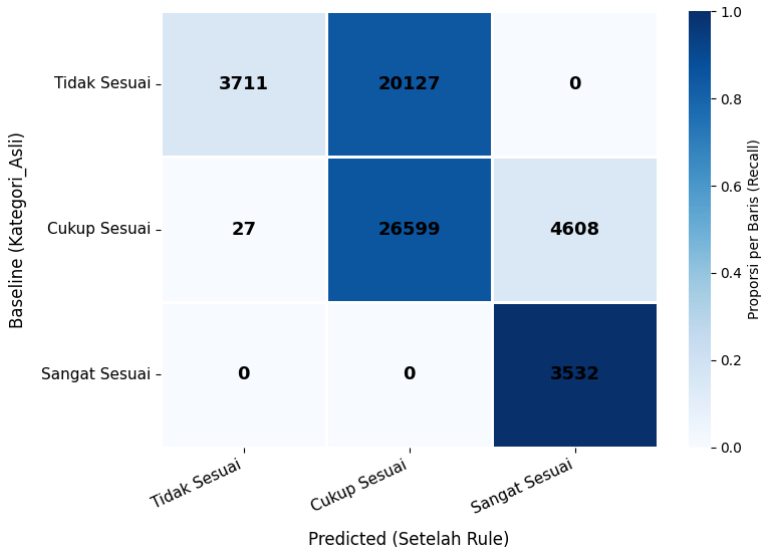
Secara keseluruhan, analisis ini menunjukkan bahwa *rule-based validation* tidak hanya mengubah distribusi kategori, tetapi juga memperbaiki interpretasi kesesuaian teks dengan mempertimbangkan aspek semantik dan konteks yang lebih luas.

4.2 Pembahasan

Pada bagian ini dilakukan pembahasan terhadap hasil penelitian yang telah diperoleh pada sub bab sebelumnya. Pembahasan difokuskan pada interpretasi hasil analisis, evaluasi performa model, serta identifikasi faktor-faktor yang memengaruhi tingkat kesesuaian teks. Selain itu, bagian ini juga membahas keterbatasan metode yang digunakan serta implikasi dari hasil penelitian dalam konteks analisis teks budaya.

4.2.1 Evaluasi Model

Pada bagian ini dilakukan evaluasi terhadap performa model yang digunakan dalam penelitian, yaitu kombinasi antara metode TF-IDF dan *rule-based validation*. Evaluasi bertujuan untuk melihat sejauh mana penerapan *rule-based* memberikan perubahan terhadap hasil klasifikasi dibandingkan dengan *baseline* TF-IDF. Perlu diperhatikan bahwa dalam penelitian ini tidak digunakan *ground truth* manual, sehingga evaluasi dilakukan dengan menjadikan hasil TF-IDF sebagai *baseline*. Oleh karena itu, nilai evaluasi yang diperoleh merepresentasikan tingkat perubahan hasil klasifikasi, bukan tingkat kebenaran absolut.



Gambar 4. 4 Matriks Perubahan Kategori *Baseline* TF-IDF dan *Rule-based*

Berdasarkan matriks perubahan kategori pada Gambar 4.4, terlihat bahwa terjadi perubahan klasifikasi yang cukup signifikan setelah penerapan *rule-based* validation. Perubahan yang paling dominan adalah perpindahan data dari kategori Tidak Sesuai ke Cukup Sesuai, serta peningkatan jumlah data pada kategori Sangat Sesuai. Pola ini menunjukkan bahwa *rule-based* validation mampu mendeteksi keterkaitan kontekstual yang tidak teridentifikasi oleh metode TF-IDF, terutama pada data yang memiliki hubungan makna namun tidak memiliki kesamaan kata secara langsung. Dengan demikian, matriks perubahan kategori dalam penelitian ini tidak hanya digunakan untuk melihat distribusi perubahan klasifikasi, tetapi juga untuk menggambarkan pengaruh penerapan *rule-based* terhadap hasil evaluasi kesesuaian teks.

Berdasarkan hasil analisis, sebagian besar perubahan klasifikasi terjadi pada data dengan tingkat kesesuaian rendah hingga sedang pada *baseline* TF-IDF. Setelah penerapan *rule-based*, banyak data berpindah ke kategori yang lebih tinggi

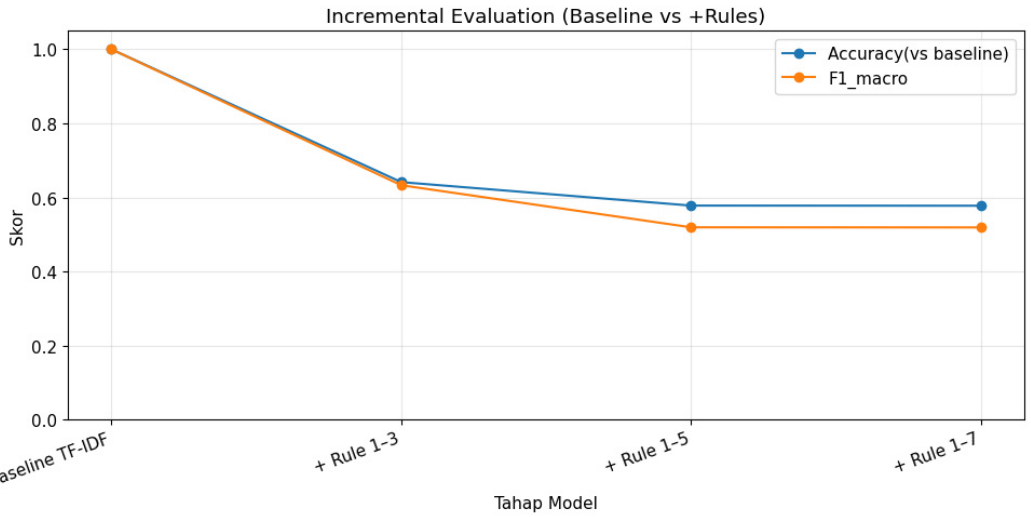
karena sistem mampu mengenali indikator kontekstual tambahan, seperti istilah budaya, pola frasa deskriptif, serta keterkaitan makna yang tidak terdeteksi melalui kesamaan kata secara langsung. Hal ini menunjukkan bahwa pendekatan *rule-based* mampu memperluas proses evaluasi kesesuaian teks dengan mempertimbangkan konteks semantik secara lebih fleksibel.

Di sisi lain, perubahan klasifikasi yang cukup besar dibandingkan *baseline* menunjukkan bahwa *rule-based* memberikan pengaruh signifikan terhadap hasil akhir evaluasi. Semakin banyak aturan yang diterapkan, semakin besar pula perubahan kategori yang dihasilkan. Kondisi ini menunjukkan adanya perbedaan pendekatan antara metode TF-IDF yang berbasis frekuensi kata dan *rule-based* validation yang mempertimbangkan konteks linguistik serta pola deskriptif budaya. Oleh karena itu, kombinasi TF-IDF dan *rule-based* dalam penelitian ini dinilai mampu menghasilkan evaluasi kesesuaian teks yang lebih kontekstual dan representatif terhadap data budaya digital.

Untuk menganalisis pengaruh masing-masing *rule* terhadap performa model, dilakukan evaluasi secara bertahap dengan menambahkan *rule* secara incremental. Hasil evaluasi ditampilkan pada Tabel 4.13 berikut.

Tabel 4. 13 Hasil *Incremental Evaluation*

Model	**Konsistensi terhadap *Baseline***	**Persentase Perubahan Klasifikasi**
Baseline TF-IDF	100%	0%
+ *Rule* 1–3	57,34%	42,66%
+ *Rule* 1–5	50,82%	49,18%
+ *Rule* 1–7	50,77%	49,23%



Gambar 4. 5 *Incremental Evaluation Rule Based*

Jika divisualisasikan dalam bentuk grafik garis pada Gambar 4.5, terlihat bahwa penambahan *rule* pada tahap awal memberikan perubahan klasifikasi yang cukup signifikan dibandingkan *baseline* TF-IDF. Hal ini menunjukkan bahwa *rule* dasar mampu membantu sistem dalam mendeteksi keterkaitan kontekstual yang sebelumnya tidak teridentifikasi oleh metode TF-IDF. Seiring dengan penambahan *rule* berikutnya, perubahan klasifikasi terhadap *baseline* menjadi semakin besar. Kondisi ini menunjukkan bahwa model menjadi lebih fleksibel dalam melakukan evaluasi kesesuaian teks, terutama pada data yang memiliki hubungan makna namun tidak memiliki kesamaan kata secara langsung.

Perubahan klasifikasi yang semakin besar juga menunjukkan adanya perbedaan pendekatan antara metode TF-IDF dan *rule-based* validation. TF-IDF berfokus pada kesamaan kata berdasarkan frekuensi kemunculan istilah, sedangkan *rule-based* mempertimbangkan pola linguistik, konteks budaya, serta frasa

deskriptif tertentu. Oleh karena itu, semakin banyak *rule* yang diterapkan, semakin besar pula pengaruh kontekstual yang diberikan terhadap hasil klasifikasi.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa penerapan *rule-based* validation memberikan dampak signifikan terhadap perubahan hasil klasifikasi. Sistem menjadi lebih sensitif dalam mendeteksi keterkaitan makna pada teks budaya, terutama pada data yang tidak teridentifikasi oleh metode TF-IDF. Dengan demikian, kombinasi antara TF-IDF dan *rule-based* dalam penelitian ini dinilai mampu menghasilkan analisis kesesuaian teks yang lebih kontekstual dan representatif terhadap data budaya digital.

Selain perubahan distribusi klasifikasi, aspek efisiensi komputasi juga menjadi bagian penting dalam evaluasi *pipeline* yang dibangun. Pengukuran waktu dilakukan pada setiap tahap secara terpisah menggunakan fungsi *time* pada Python, sehingga kontribusi masing-masing tahap terhadap total waktu dapat diidentifikasi secara akurat. Hal ini penting untuk memahami di mana letak *bottleneck* dalam sistem dan sejauh mana penambahan komponen *rule-based* memengaruhi beban komputasi secara keseluruhan. Berdasarkan hasil eksperimen terhadap 58.604 entri data, total waktu pemrosesan keseluruhan *pipeline* mencapai 5 jam 28 menit 1 detik (19.681,02 detik). Rincian waktu per tahap dapat dilihat pada Tabel 4.14 berikut.

Tabel 4. 14 Waktu Pemrosesan *Pipeline* Analisis Kesesuaian Teks

Tahap	**Waktu (Detik)**	**Waktu (Menit)**	***Throughput* (Entri/Detik)**
Data Preparation (Cleaning + Stemming)	19.572,55	326,21	3,0
TF-IDF + *Cosine Similarity*	59,00	0,98	993,3
Rule-Based Validation	26,18	0,44	2.238,6
Incremental Evaluation	23,28	0,39	2.516,9
Total	**19.681,02**	**328,02**	—

Berdasarkan tabel tersebut, terlihat bahwa sebagian besar waktu pemrosesan didominasi oleh tahap *Data Preparation*, yaitu sekitar 19.572,55 detik atau 99,5% dari total waktu keseluruhan. Hal ini disebabkan oleh proses *stemming* yang dilakukan secara sekuensial terhadap seluruh 58.604 entri menggunakan *library* Sastrawi, di mana setiap entri memerlukan analisis morfologi bahasa Indonesia yang cukup intensif secara komputasi. Rata-rata throughput pada tahap ini hanya mencapai 3 entri per detik.

Sebaliknya, tahap TF-IDF + *Cosine Similarity*, *Rule-Based Validation*, dan *Incremental Evaluation* menunjukkan efisiensi yang jauh lebih tinggi. Tahap TF-IDF mampu memproses sekitar 993 entri per detik, sementara *Rule-Based Validation* mencapai 2.238 entri per detik dan *Incremental Evaluation* mencapai 2.517 entri per detik. Perbedaan *throughput* yang signifikan antara tahap *Data Preparation* dan tahap-tahap berikutnya menunjukkan bahwa *bottleneck* utama dalam *pipeline* ini terletak pada proses *stemming*, bukan pada perhitungan TF-IDF maupun penerapan *rule-based*. Dengan demikian, penambahan *rule-based validation* tidak memberikan overhead waktu yang signifikan hanya menambah sekitar 26 detik dari total keseluruhan sehingga dari segi efisiensi, pendekatan ini dinilai layak diterapkan pada dataset berskala besar.

Untuk menguji kemampuan sistem secara lebih terperinci, dilakukan rancangan percobaan menggunakan 15 data uji yang mencakup 5 tema budaya, yaitu Tari Tradisional, Masakan Tradisional, Lagu Daerah, Situs Budaya, dan Cerita Rakyat. Pemilihan kelima tema tersebut didasarkan pada representasi kategori konten yang paling banyak ditemukan dalam dataset budaya-

indonesia.org, sehingga hasil pengujian diharapkan dapat mencerminkan kondisi data secara lebih representatif. Setiap tema diuji menggunakan 3 variasi deskripsi yang dirancang dengan tingkat kespesifikan yang berbeda-beda, mulai dari deskripsi yang menyebutkan nama objek secara langsung hingga deskripsi yang bersifat generik tanpa menyebut nama objek sama sekali. Rancangan variasi ini bertujuan untuk mensimulasikan kondisi nyata di mana kualitas deskripsi antar konten pada platform *crowdsourcing* sangat beragam, sebagaimana ditunjukkan pada Tabel 4.15 berikut.

Tabel 4. 15 Rancangan Tingkat Variasi Data Uji

No	Level	Ciri Utama
1	Data Uji 1	Nama objek/judul disebut langsung dalam deskripsi
2	Data Uji 2	Nama objek tidak disebut langsung, namun ciri khas masih kuat
3	Data Uji 3	Nama objek tidak disebut dan deskripsi dibuat lebih umum

Rancangan pada Tabel 4.14 bertujuan untuk menguji sejauh mana sistem mampu mengenali kesesuaian antara judul dan deskripsi pada kondisi yang semakin menantang, dari deskripsi yang sangat eksplisit hingga deskripsi yang bersifat generik tanpa menyebutkan nama objek secara langsung. Hasil pengujian sistem menggunakan kombinasi TF-IDF dan *rule-based validation* terhadap 15 skenario menggunakan 3 data uji tersebut disajikan pada Tabel 4.16 berikut.

Tabel 4. 16 Hasil Rancangan 15 Skenario Percobaan

No	Tema	Data Uji	Judul	Deskripsi	Hasil
1	Tari Tradisional	Data Uji 1	Tari Kecak Bali	Tari Kecak merupakan tarian tradisional yang berasal dari Bali dan sering dipentaskan dalam acara budaya maupun pertunjukan wisata. Tarian ini dimainkan secara berkelompok dengan iringan suara khas dari para penarinya tanpa	TF-IDF = 0.410 kategori cukup sesuai + rule based = kategori sangat sesuai dengan rule detail = Title word overlap,

No	Tema	Data Uji	Judul	Deskripsi	Hasil
				menggunakan alat musik utama. Tari Kecak menjadi salah satu kesenian daerah yang terkenal dan banyak dikenal oleh wisatawan lokal maupun mancanegara.	cultural context, synonym detetcted, descriptive phrase
2	Tari Tradisional	Data Uji 2	Tari Kecak Bali	Tarian tradisional sering dipentaskan dalam berbagai acara budaya dan pertunjukan seni daerah di Indonesia. Kesenian ini biasanya dimainkan secara berkelompok dan menjadi salah satu hiburan budaya yang masih dilestarikan oleh masyarakat. Pertunjukan budaya tersebut juga sering menarik perhatian wisatawan yang datang untuk melihat kesenian tradisional daerah.	TF-IDF = 0.085 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = Title word overlap, synonym detetcted, descriptive phrase
3	Tari Tradisional	Data Uji 3	Tari Kecak Bali	Kesenian tradisional dari Indonesia sering dimainkan secara berkelompok dalam berbagai pertunjukan seni daerah dan acara wisata. Kesenian tersebut menggunakan iringan suara khas dari para pemain dan menjadi salah satu hiburan yang cukup terkenal di kalangan wisatawan lokal maupun mancanegara.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context
4	Masakan Tradisional	Data Uji 1	Rendang Padang	Rendang merupakan makanan tradisional khas Padang yang dimasak menggunakan santan dan berbagai rempah-rempah pilihan. Proses memasaknya membutuhkan waktu cukup lama agar bumbu dapat meresap dengan sempurna ke dalam daging. Rendang dikenal memiliki cita rasa gurih dan pedas serta sering disajikan dalam acara adat dan kegiatan keluarga.	TF-IDF = 0.163 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = culinary relevance, Title word overlap, cultural context, descriptive phrase
5	Masakan Tradisional	Data Uji 2	Rendang Padang	Masakan tradisional Padang Indonesia dikenal memiliki rasa yang kuat dan menggunakan banyak rempah dalam proses pengolahannya. Kuliner daerah biasanya disajikan dalam acara keluarga maupun kegiatan budaya masyarakat. Beberapa makanan khas juga menjadi	TF-IDF = 0.182 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = Title word overlap, cultural context, descriptive phrase

No	Tema	Data Uji	Judul	Deskripsi	Hasil
				bagian penting dalam tradisi masyarakat setempat.	
6	Masakan Tradisional	Data Uji 3	Rendang Padang	Daging sapi dipotong menjadi beberapa bagian lalu dimasak menggunakan santan bersama bawang merah, bawang putih, cabai, lengkuas, jahe, serai, dan daun jeruk. Semua bahan dimasak dengan api kecil dalam waktu yang cukup lama hingga kuah mengering dan bumbu meresap sempurna ke dalam daging.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = culinary relevance
7	Lagu Daerah	Data Uji 1	Ampar-Ampar Pisang	Ampar-Ampar Pisang merupakan lagu daerah yang sering dinyanyikan dalam kegiatan budaya maupun pembelajaran seni tradisional di sekolah. Lagu ini memiliki irama khas yang mudah dikenali dan menjadi salah satu bagian dari kesenian daerah Indonesia. Lagu tradisional tersebut masih sering diperkenalkan kepada generasi muda sebagai bentuk pelestarian budaya.	TF-IDF = 0.452 kategori cukup sesuai. + rule based = kategori sangat sesuai dengan rule detail = Title word overlap, cultural context, descriptive phrase
8	Lagu Daerah	Data Uji 2	Ampar-Ampar Pisang	Lagu daerah Indonesia memiliki irama khas dan sering digunakan dalam kegiatan budaya serta pertunjukan seni masyarakat. Musik daerah biasanya diperkenalkan melalui pembelajaran budaya maupun festival kesenian daerah. Kesenian musik tradisional masih menjadi bagian penting dalam pelestarian budaya Indonesia.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context, descriptive phrase
9	Lagu Daerah	Data Uji 3	Ampar-Ampar Pisang	Lagu daerah Indonesia memiliki irama yang khas dan sering dinyanyikan dalam kegiatan seni maupun pembelajaran di sekolah. Musik tradisional tersebut menjadi salah satu bagian penting dalam pelestarian kesenian daerah dan masih dikenalkan kepada generasi muda hingga saat ini.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context
10	Situs Budaya	Data Uji 1	Candi Borobudur	Candi Borobudur merupakan situs budaya bersejarah yang menjadi salah satu peninggalan Buddha terbesar di Indonesia. Bangunan ini memiliki banyak	TF-IDF = 0.329 kategori cukup sesuai + rule based = kategori sangat

No	Tema	Data Uji	Judul	Deskripsi	Hasil
				relief yang menggambarkan kehidupan masyarakat pada masa lampau. Candi Borobudur juga menjadi salah satu destinasi wisata budaya yang sering dikunjungi wisatawan lokal maupun mancanegara.	sesuai dengan rule detail = Title word overlap, cultural context, descriptive phrase
11	Situs Budaya	Data Uji 2	Candi Borobudur	Situs budaya sering dikunjungi wisatawan karena memiliki nilai sejarah dan arsitektur yang khas bagi budaya Indonesia. Bangunan peninggalan sejarah biasanya menjadi bagian penting dalam pelestarian budaya daerah. Tempat bersejarah seperti candi juga sering digunakan sebagai sarana pembelajaran budaya dan sejarah masyarakat.	TF-IDF = 0.145 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = Title word overlap, cultural context, descriptive phrase
12	Situs Budaya	Data Uji 3	Candi Borobudur	Bangunan bersejarah di Indonesia memiliki relief dan arsitektur khas yang menggambarkan kehidupan masyarakat pada masa lampau. Tempat tersebut menjadi salah satu destinasi wisata yang banyak dikunjungi karena memiliki nilai sejarah dan keunikan bentuk bangunannya.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context, descriptive phrase
13	Cerita Rakyat	Data Uji 1	Malin Kundang	Malin Kundang merupakan cerita rakyat yang menceritakan kisah seorang anak yang durhaka kepada ibunya dan menjadi pelajaran moral bagi masyarakat. Cerita rakyat ini sering diperkenalkan kepada anak-anak melalui pembelajaran budaya maupun pertunjukan seni daerah. Kisah tersebut menjadi salah satu cerita tradisional yang cukup terkenal di Indonesia.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = Title word overlap, cultural context, descriptive phrase
14	Cerita Rakyat	Data Uji 2	Malin Kundang	Cerita tradisional Indonesia banyak mengandung pesan moral dan masih dikenalkan kepada masyarakat melalui pembelajaran budaya. Kisah rakyat biasanya menceritakan kehidupan masyarakat serta nilai-nilai yang diwariskan secara turun-temurun. Cerita daerah juga menjadi bagian penting dalam pelestarian budaya Indonesia.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context

No	Tema	Data Uji	Judul	Deskripsi	Hasil
15	Cerita Rakyat	Data Uji 3	Malin Kundang	Cerita rakyat Indonesia sering menceritakan kehidupan keluarga dan hubungan antara orang tua dengan anak dalam kehidupan sehari-hari. Kisah tersebut biasanya mengandung pesan moral yang diwariskan secara turun-temurun kepada masyarakat melalui pembelajaran dan pertunjukan seni daerah.	TF-IDF = 0.000 kategori tidak sesuai + rule based = kategori cukup sesuai dengan rule detail = cultural context

Berdasarkan hasil percobaan pada Tabel 4.15, terdapat beberapa pola yang dapat diidentifikasi. Pada Data Uji 1, tiga dari lima tema yaitu Tari Tradisional, Lagu Daerah, dan Situs Budaya berhasil mencapai kategori Sangat Sesuai setelah penerapan rule-based, sementara dua tema lainnya yaitu Masakan Tradisional dan Cerita Rakyat tetap berada pada kategori Cukup Sesuai. Hal ini menunjukkan bahwa ketika nama objek disebut langsung dalam deskripsi, sistem mampu mengenali kesesuaian secara optimal melalui kombinasi skor TF-IDF dan *rule title word overlap*, meskipun tidak semua tema mencapai hasil tertinggi karena keterbatasan representasi kosakata dalam bobot TF-IDF.

Pada Data Uji 2 dan Data Uji 3, seluruh tema menghasilkan kategori Cukup Sesuai setelah penerapan rule-based, meskipun skor TF-IDF-nya sangat rendah bahkan mencapai 0,000. Kondisi ini menunjukkan bahwa *rule-based* mampu mendeteksi keterkaitan kontekstual melalui pengenalan pola budaya (*cultural context*) dan frasa deskriptif (*descriptive phrase*) yang tidak tertangkap oleh pendekatan berbasis frekuensi kata. Pengaruh *rule-based* terhadap hasil akhir sistem dapat dilihat secara jelas pada Tabel 4.17 dan Tabel 4.18 berikut.

Tabel 4. 17 Rekap hasil per level data uji

Level	**TF-IDF Saja**	**%**	**TF-IDF+ *Rule-based***	**%**	**Perubahan**
Data Uji 1 (nama disebut langsung)	Cukup Sesuai (3×)	60%	Sangat Sesuai (3×)	60%	CS → SS
	Tidak Sesuai (2×)	40%	Cukup Sesuai (2×)	40%	TS → CS
Data Uji 2 (ciri khas kuat, tanpa nama)	Tidak Sesuai (5×)	100%	Cukup Sesuai (5×)	100%	TS → CS
Data Uji 3 (deskripsi generik, tanpa nama)	Tidak Sesuai (5×)	100%	Cukup Sesuai (5×)	100%	TS → CS

Tabel 4. 18 Rekap Hasil per Tema

Tema	**Hasil TF-IDF saja**			**Hasil TF-IDF + Rule-based**		
	Uji 1	**Uji 2**	**Uji 3**	**Uji 1**	**Uji 2**	**Uji 3**
Tari Tradisional	CS	TS	TS	SS	CS	CS
Masakan Tradisional	TS	TS	TS	CS	CS	CS
Lagu Daerah	CS	TS	TS	SS	CS	CS
Situs Budaya	CS	TS	TS	SS	CS	CS
Cerita Rakyat	TS	TS	TS	CS	CS	CS

Keterangan:
SS = Sangat Sesuai,
CS = Cukup Sesuai,
TS = Tidak Sesuai

Berdasarkan Tabel 4.17 dan Tabel 4.18 di atas secara konsisten memperlihatkan pengaruh signifikan *rule-based* terhadap hasil akhir sistem. Tanpa perlakuan *rule-based,* TF-IDF saja hanya mampu mengklasifikasikan 3 dari 15 skenario sebagai Cukup Sesuai dan sama sekali tidak ada yang mencapai Sangat Sesuai, sementara 12 skenario lainnya terklasifikasi ke dalam kategori Tidak Sesuai. Setelah perlakuan *rule-based* diterapkan, seluruh 15 skenario mengalami peningkatan kategori tidak ada satu pun yang tersisa di kategori Tidak Sesuai dengan 3 skenario mencapai Sangat Sesuai (20%) dan 12 skenario berada pada

kategori Cukup Sesuai (80%). Peningkatan ini terjadi secara merata di seluruh tema dan semua level data uji, yang menunjukkan bahwa *rule-based* bukan sekadar pelengkap, melainkan komponen yang secara nyata menentukan kemampuan sistem dalam mengenali kesesuaian kontekstual antara judul dan deskripsi konten budaya.

4.2.2 Analisis Kesalahan

Meskipun penerapan metode TF-IDF yang dikombinasikan dengan *rule-based validation* mampu meningkatkan hasil klasifikasi secara signifikan, masih terdapat sejumlah kasus di mana sistem menghasilkan klasifikasi yang kurang tepat. Oleh karena itu, dilakukan analisis kesalahan (*error analysis*) untuk mengidentifikasi pola kegagalan sistem serta memahami faktor-faktor yang memengaruhi ketidaktepatan hasil.

Analisis ini dilakukan dengan mengamati beberapa sampel data yang masih dikategorikan tidak sesuai atau mengalami kesalahan klasifikasi setelah penerapan *rule-based*. Beberapa contoh data yang masih dikategorikan sebagai *Tidak Sesuai* ditampilkan pada Tabel 4.19 berikut.

Tabel 4. 19 Contoh Data Tidak Sesuai

No	Judul	Deskripsi	Kategori
1	Test2	Begono (sumber: E-b ook Mahakarya 5000 Resep Makanan dan Minuman di Indonesia)	Tidak Sesuai
2	Bebotok	Jawa tengah terkenal akan budayanya	Tidak Sesuai
3	MERSIBEREN TANDA BURJU	Dalam tahap ini peranan pihak ketiga tetap penting.dari pihak sigadis sebagai saksinya adalah bibinya sedangkan dari pihak laki adalah sini nana(satu marga).Pada saat pertukaran cincin dilakukan petukaran barang(cincin,kain dan lain2) kadang2 diakhiri dengan membuat ikrar atau janji yang disebut merbulaban. Kemudian salah satu pengetuai(saksi) mengucapkan kata2;kongpe uratni buluh,kongen deng urat ni teladan,kong pe katani hukum	Tidak Sesuai

		kongen deng kata ni padan Artinya walaupun hokum memiliki kekuatan namun lebih kuat lagi perjanjian sumber :	
4	Padungku	(Ralat)	Tidak Sesuai
5	Adat Pernikahan Betawi	asdfghjkl	Tidak Sesuai

Berdasarkan contoh pada Tabel 4.19, dapat disimpulkan bahwa data yang dikategorikan sebagai *Tidak Sesuai* diduga disebabkan oleh kualitas kesesuaian dan keterkaitan antara *Judul* dan *Deskripsi*. Beberapa permasalahan yang ditemukan meliputi judul yang tidak representatif, deskripsi yang terlalu umum atau tidak relevan, penggunaan bahasa yang sulit dikenali oleh sistem, serta adanya *noise* seperti teks tidak bermakna atau deskripsi kosong. Kondisi-kondisi tersebut menyebabkan nilai kemiripan menjadi rendah sehingga sistem mengklasifikasikan data sebagai *Tidak Sesuai*. Hal ini menunjukkan bahwa selain metode yang digunakan, kualitas data sangat berpengaruh terhadap hasil klasifikasi.

Berdasarkan hasil analisis terhadap data yang masuk di kategori tidak sesuai, terdapat beberapa faktor yang diduga menyebabkan sistem tidak dapat mengklasifikasikan kesesuaian teks secara akurat.

1. **Deskripsi Kosong atau Sangat Pendek**

   Salah satu penyebab utama kesalahan adalah adanya deskripsi yang kosong atau terlalu singkat. Pada kondisi ini, sistem tidak memiliki cukup informasi untuk melakukan perbandingan dengan judul. Selain itu, *rule-based* juga memberikan penalti pada deskripsi yang terlalu pendek, sehingga semakin memperkuat kemungkinan data dikategorikan sebagai *Tidak Sesuai*.

2. Deskripsi Tidak Relevan dengan Judul

Kesalahan juga terjadi pada data di mana deskripsi tidak memiliki keterkaitan dengan judul. Hal ini sering ditemukan pada data hasil *crowdsourcing*, di mana pengguna dapat mengisi deskripsi secara bebas tanpa kontrol kualitas yang ketat. Sebagai contoh, sebuah judul yang merujuk pada artefak budaya tertentu dapat memiliki deskripsi yang membahas hal lain yang tidak berkaitan. Dalam kondisi ini, baik TF-IDF maupun *rule-based* tidak dapat menemukan indikator kesesuaian yang cukup kuat.

3. Tidak Terdapat Overlap Kata

Metode TF-IDF sangat bergantung pada kemunculan kata yang sama antara judul dan deskripsi. Oleh karena itu, apabila tidak terdapat *overlap* kata, nilai similarity yang dihasilkan akan sangat rendah, bahkan mendekati nol.

Meskipun *rule-based* telah menambahkan mekanisme deteksi konteks dan sinonim, dalam beberapa kasus perbedaan istilah yang terlalu jauh tetap menyebabkan sistem gagal mengenali kesesuaian.

4. Variasi Bahasa dan Istilah Daerah

Dalam beberapa kasus, perbedaan bahasa atau penggunaan istilah daerah menyebabkan sistem tidak dapat mengenali kesamaan makna antara judul dan deskripsi. Hal ini menjadi tantangan tersendiri karena TF-IDF tidak mampu memahami hubungan semantik di luar kesamaan kata secara langsung.

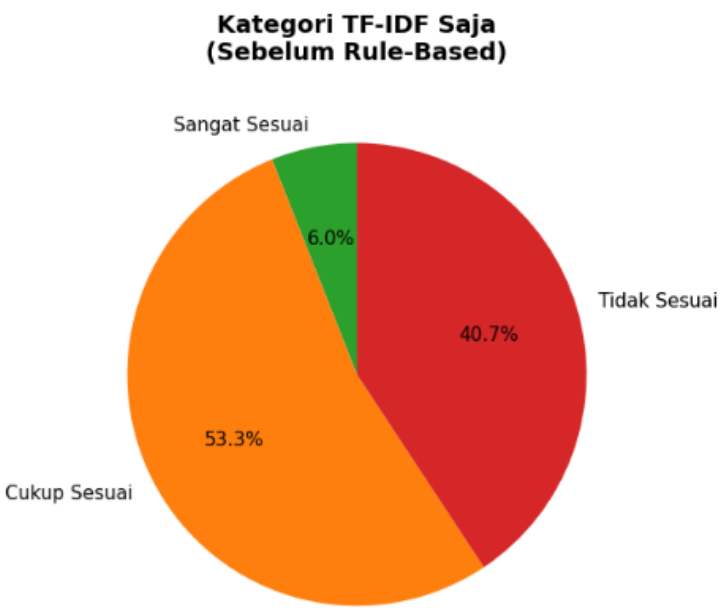
Jika diamati secara keseluruhan, jumlah data yang mengalami kesalahan klasifikasi relatif lebih kecil dibandingkan dengan jumlah data yang berhasil diperbaiki oleh *rule-based* validation. Hal ini menunjukkan bahwa meskipun masih terdapat keterbatasan, pendekatan yang digunakan dalam penelitian ini tetap memberikan peningkatan yang signifikan terhadap kualitas hasil klasifikasi.

Selain itu, sebagian besar kesalahan yang terjadi disebabkan oleh kualitas data yang rendah, bukan semata-mata karena kelemahan metode. Dengan demikian, peningkatan kualitas data, seperti pengisian deskripsi yang lebih lengkap dan relevan, berpotensi meningkatkan performa sistem secara keseluruhan.

4.2.3 Pembahasan dan Integrasi Islam

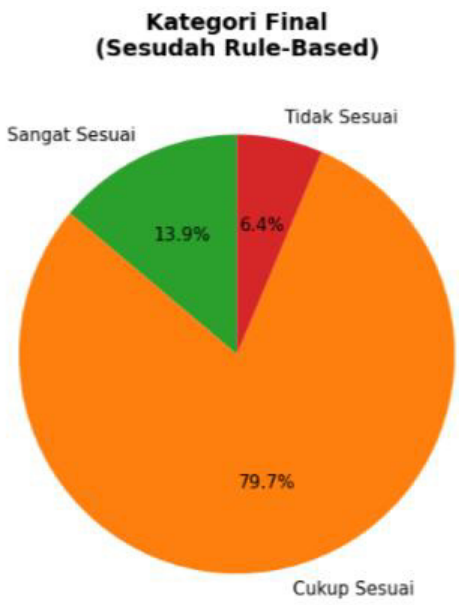
Pada bagian ini dilakukan pembahasan secara menyeluruh terhadap hasil penelitian yang telah diperoleh. Pembahasan difokuskan pada tiga aspek utama, yaitu perbandingan antara metode *baseline* TF-IDF dan *rule-based validation*, implementasi sistem dalam bentuk aplikasi berbasis web, serta integrasi nilai-nilai Islam dalam proses analisis informasi.

Perbandingan hasil antara metode *baseline* TF-IDF dan metode yang telah dikombinasikan dengan *rule-based validation* divisualisasikan dalam bentuk diagram lingkaran (*pie chart*) yang ditunjukkan pada Gambar 4.6 dan Gambar 4.7.



Gambar 4. 6 Kategori TF-IDF *Baseline*

Pada Gambar 4.6 (*baseline* TF-IDF), terlihat bahwa distribusi kategori didominasi oleh dua kategori utama, yaitu *Cukup Sesuai* dan *Tidak Sesuai*. Kategori *Sangat Sesuai* hanya memiliki proporsi yang sangat kecil dibandingkan kategori lainnya. Hal ini menunjukkan bahwa metode TF-IDF cenderung menghasilkan nilai kesesuaian pada tingkat menengah, yang disebabkan oleh pendekatan yang hanya mempertimbangkan frekuensi kemunculan kata tanpa memahami konteks.



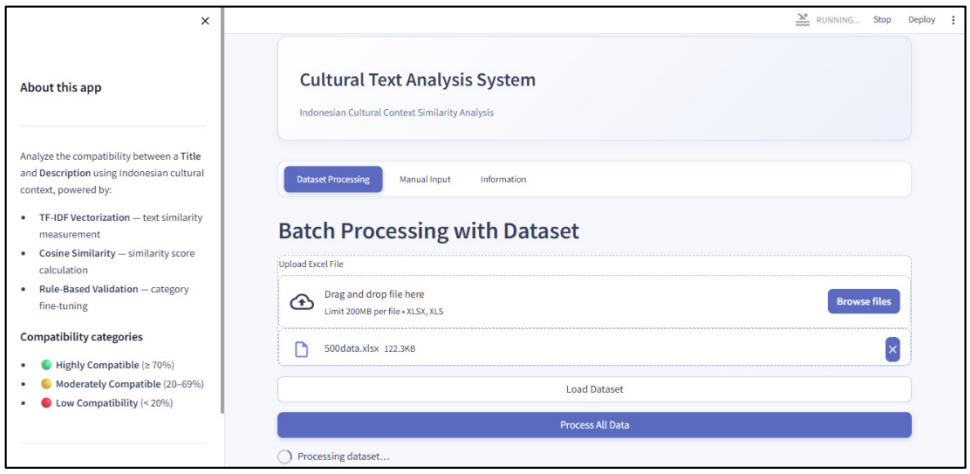
Gambar 4. 7 Kategori Setelah *Rule Based*

Sebaliknya, pada Gambar 4.7 (setelah *rule-based*), terlihat perubahan distribusi yang cukup signifikan. Proporsi kategori *Tidak Sesuai* mengalami penurunan yang drastis, sementara kategori *Cukup Sesuai* dan *Sangat Sesuai* mengalami peningkatan yang cukup besar. Secara visual, bagian pie untuk *Tidak Sesuai* yang sebelumnya mendominasi menjadi jauh lebih kecil, sedangkan *Cukup Sesuai* menjadi kategori yang paling dominan, diikuti oleh peningkatan signifikan pada *Sangat Sesuai*. Perubahan ini menunjukkan bahwa *rule-based validation* mampu memperbaiki hasil klasifikasi dengan mengidentifikasi kesesuaian yang sebelumnya tidak terdeteksi oleh TF-IDF. Hal ini terutama terjadi pada data yang memiliki keterkaitan makna namun tidak memiliki kesamaan kata secara langsung. Jika diamati lebih lanjut, peningkatan pada kategori *Sangat Sesuai* menunjukkan bahwa *rule-based* tidak hanya memperbaiki data yang sebelumnya dianggap tidak sesuai, tetapi juga mampu meningkatkan tingkat keyakinan model terhadap data

yang sudah cukup sesuai. Dengan kata lain, *rule-based* berperan dalam memperhalus klasifikasi sehingga lebih mendekati kondisi sebenarnya. Selain itu, perubahan distribusi ini juga sejalan dengan hasil evaluasi model yang menunjukkan bahwa nilai recall meningkat. Hal ini mengindikasikan bahwa sistem menjadi lebih sensitif dalam mendeteksi kesesuaian, meskipun dengan konsekuensi meningkatnya perbedaan terhadap *baseline*.

Hasil penelitian ini tidak hanya berhenti pada pengembangan metode, tetapi juga diimplementasikan dalam bentuk sistem berbasis web yang ditujukan untuk mempermudah pengguna dalam melakukan analisis kesesuaian teks secara praktis.

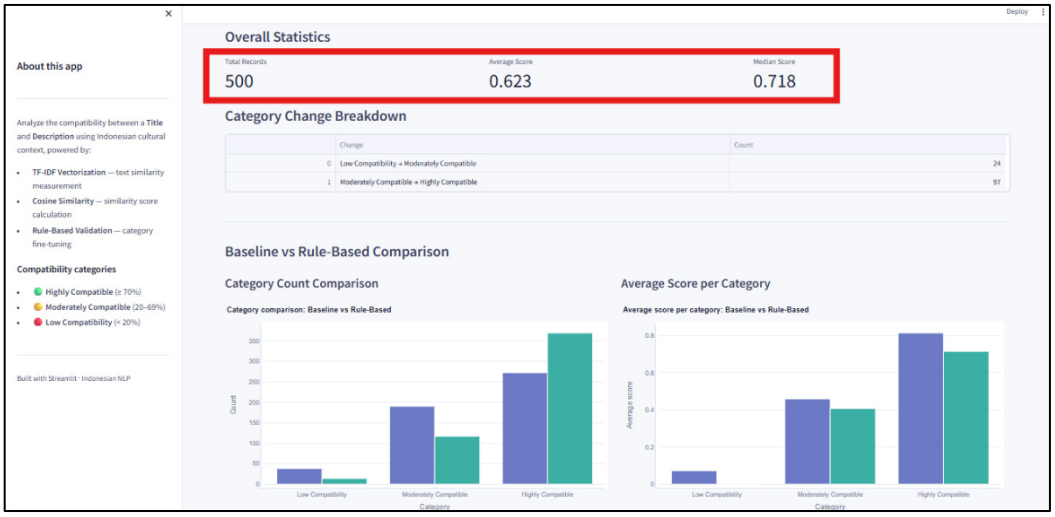
Tampilan antarmuka sistem ditunjukkan pada Gambar 4.8.



Gambar 4. 8 Screenshot Tampilan Untuk Proses Dataset

Berdasarkan tampilan pada Gambar 4.8, sistem menyediakan fitur *Batch Processing with Dataset*, di mana pengguna dapat mengunggah file Excel yang berisi data judul dan deskripsi. Setelah dataset dimuat, pengguna dapat menjalankan proses analisis dengan menekan tombol *Process All Data*, kemudian sistem akan secara otomatis melakukan serangkaian proses yang meliputi preprocessing teks,

pembentukan vektor TF-IDF, perhitungan *cosine similarity*, serta penerapan *rule-based validation*. Hasil analisis ditampilkan pada bagian *Processing Results* dalam bentuk tabel yang memuat nilai skor TF-IDF, kategori sebelum dan sesudah *rule-based*, serta detail *rule* yang terpicu pada setiap data.



Gambar 4. 9 Screenshot Tampilan Visualisasi Hasil Proses Dataset

Selanjutnya, Gambar 4.9 pada bagian *Overall Statistics*, sistem menyajikan informasi agregat berupa jumlah data yang diproses (500 entri), rata-rata skor (0,623), serta median skor (0,718). Selain itu, bagian *Category Change Breakdown* menampilkan perubahan kategori yang terjadi.

Gambar 4. 10 Visualisasi Grafik Perbandingan

Berdasarkan Gambar 4.10, sistem juga dilengkapi dengan visualisasi tambahan berupa grafik batang yang memperlihatkan perbandingan jumlah kategori sebelum dan sesudah *rule-based*, serta rata-rata skor pada setiap kategori. Implementasi ini menunjukkan bahwa sistem tidak hanya mampu melakukan analisis secara akurat, tetapi juga menyajikan hasil secara visual dan informatif, sehingga pengguna dapat dengan mudah memahami perubahan hasil tanpa harus memahami proses teknis yang kompleks.

Penelitian ini juga memiliki relevansi dengan nilai-nilai Islam, khususnya dalam hal verifikasi informasi. Prinsip ini sejalan dengan ajaran dalam Al-Hujurat ayat 6 yang menekankan pentingnya melakukan klarifikasi terhadap informasi sebelum mempercayai atau menyebarkannya.

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ۝٦

"Wahai orang-orang yang beriman, apabila datang kepadamu seorang fasik membawa suatu berita, maka periksalah dengan teliti (tabayyun), agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui

keadaannya, yang menyebabkan kamu menyesal atas perbuatanmu itu." (QS. Al-Hujurat [49]: 6)

Menurut M. Quraish Shihab dalam *Tafsir Al-Misbah*, ayat ini mengandung perintah untuk melakukan pemeriksaan dan klarifikasi terhadap suatu informasi sebelum dijadikan dasar dalam mengambil keputusan. Tujuannya adalah agar seseorang tidak terburu-buru dalam memberikan penilaian yang dapat menimbulkan kesalahan, kerugian, maupun penyesalan di kemudian hari.

Dalam ayat tersebut dijelaskan bahwa apabila seseorang menerima suatu berita, maka berita tersebut harus diperiksa terlebih dahulu kebenarannya agar tidak menimbulkan kesalahan atau kerugian. Prinsip ini sangat relevan dengan konteks penelitian, di mana sistem dirancang untuk menganalisis apakah suatu deskripsi benar-benar sesuai dengan judul yang diberikan. Dengan demikian, proses verifikasi kesesuaian informasi yang dilakukan menggunakan metode TF-IDF dan rule-based validation sejalan dengan nilai tabayyun yang diajarkan dalam QS. Al-Hujurat ayat 6.

Jika dikaitkan dengan proses dalam sistem, metode TF-IDF dapat dianggap sebagai tahap awal dalam mengidentifikasi kesesuaian informasi secara objektif berdasarkan kemunculan kata. Sementara itu, *rule-based validation* berperan sebagai tahap lanjutan yang melakukan verifikasi lebih mendalam dengan mempertimbangkan konteks, makna, dan pola bahasa.

Dengan demikian, proses analisis dalam penelitian ini dapat dipandang sebagai bentuk implementasi prinsip tabayyun (klarifikasi informasi), di mana informasi tidak langsung diterima begitu saja, melainkan dianalisis terlebih dahulu

sebelum diambil kesimpulan. Integrasi nilai ini menunjukkan bahwa penelitian tidak hanya memiliki manfaat dari sisi teknologi, tetapi juga memiliki landasan etis yang sejalan dengan ajaran Islam dalam menjaga kebenaran informasi.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode TF-IDF mampu digunakan sebagai pendekatan awal dalam mengukur kesesuaian antara judul dan deskripsi teks budaya melalui perhitungan *cosine similarity*. Nilai similarity yang dihasilkan berada pada rentang 0 hingga 1, dengan mayoritas data 53,3% yaitu dengan total 31.234 entri berada pada kategori Cukup Sesuai, sementara 40,7% dengan total 23.838 entri terklasifikasi sebagai Tidak Sesuai dan hanya 6% dengan total 3.532 entri yang mencapai kategori Sangat Sesuai. Namun demikian, metode TF-IDF memiliki keterbatasan karena hanya mempertimbangkan kemunculan kata secara literal, sehingga belum mampu menangkap hubungan makna yang lebih dalam seperti sinonim, konteks budaya, dan pola deskriptif.

Untuk mengatasi keterbatasan tersebut, penelitian ini mengimplementasikan rule-based validation sebagai metode tambahan dengan tujuh rule. Hasil penerapan rule-based menunjukkan peningkatan yang sangat signifikan dalam distribusi kategori kesesuaian. Kategori Sangat Sesuai meningkat dari 3.532 entri (6%) menjadi 8.140 entri (13,9%), atau meningkat sebesar 130,4%. Kategori Cukup Sesuai juga meningkat dari 31.234 entri (53,3%) menjadi 46.726 entri (79,7%), atau meningkat sebesar 49,6%. Peningkatan pada kedua kategori tersebut menunjukkan bahwa penerapan rule-based validation secara signifikan

mampu meningkatkan proporsi data yang terklasifikasi sesuai, dari 59,3% menjadi 93,6% secara keseluruhan.

Dari tujuh rule yang diimplementasikan, R_overlap menjadi rule dengan frekuensi *trigger* tertinggi sebesar 54.365 kali, diikuti R_cultural sebesar 19.168 kali, yang mengindikasikan bahwa deteksi konteks budaya dan kesamaan leksikal merupakan faktor paling dominan dalam peningkatan klasifikasi.

Hasil evaluasi terhadap 15 skenario percobaan lintas lima tema budaya (tari tradisional, masakan tradisional, lagu daerah, situs budaya, dan cerita rakyat) dengan tiga level data uji menunjukkan bahwa perlakuan kombinasi TF-IDF dan *rule-based* terhadap skenario percobaan memberikan peningkatan klasifikasi yang signifikan. Tanpa perlakuan *rule-based,* TF-IDF hanya mampu mengklasifikasikan 3 dari 15 skenario (20%) ke dalam kategori Cukup Sesuai dan tidak satu pun mencapai kategori Sangat Sesuai, sedangkan 12 skenario (80%) terklasifikasi sebagai Tidak Sesuai. Setelah perlakuan *rule-based* diterapkan, seluruh 15 skenario (100%) mengalami peningkatan kategori, dengan 3 skenario (20%) mencapai Sangat Sesuai dan 12 skenario (80%) berada pada kategori Cukup Sesuai. Hasil evaluasi inkremental juga menunjukkan bahwa penambahan jumlah *rule* yang diterapkan berkorelasi dengan peningkatan perubahan klasifikasi, yang mengindikasikan adanya *trade-off* terukur antara konsistensi terhadap *baseline* dan kemampuan sistem dalam menangkap konteks semantik yang lebih luas.

Selain itu, penelitian ini memiliki relevansi dengan nilai-nilai Islam, khususnya dalam prinsip tabayyun atau verifikasi informasi sebagaimana dijelaskan dalam QS. Al-Hujurat ayat 6. Proses analisis kesesuaian yang dilakukan

sistem yakni memverifikasi relevansi suatu deskripsi terhadap judulnya sebelum menghasilkan klasifikasi akhir mencerminkan pentingnya pemeriksaan informasi secara sistematis sebelum menarik kesimpulan, sehingga penelitian ini tidak hanya berkontribusi secara teknis, tetapi juga memiliki landasan etis dalam pengolahan informasi.

Secara keseluruhan, kombinasi TF-IDF dan *rule-based validation* mampu meningkatkan kualitas klasifikasi kesesuaian teks budaya secara terukur, ditandai dengan penurunan kategori Tidak Sesuai sebesar 84,3% dan peningkatan kategori Sangat Sesuai sebesar 130,4%, sehingga menghasilkan klasifikasi yang lebih kontekstual, representatif, dan relevan dengan kebutuhan pengguna.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diajukan untuk pengembangan penelitian selanjutnya maupun implementasi sistem ke depan.

Pertama, dari sisi metode, penelitian selanjutnya disarankan untuk mengembangkan pendekatan yang digunakan dengan memanfaatkan teknik *Natural Language Processing* yang lebih canggih, seperti *word embedding* (Word2Vec, FastText) atau model berbasis *deep learning* seperti BERT. Pendekatan tersebut diharapkan mampu memahami hubungan semantik antar kata secara lebih mendalam dibandingkan metode TF-IDF, sehingga dapat meningkatkan kemampuan sistem dalam mendeteksi kesesuaian teks secara lebih akurat.

Kedua, dari sisi data, disarankan untuk meningkatkan kualitas dataset yang digunakan, terutama dengan memastikan bahwa setiap deskripsi memiliki keterkaitan yang jelas dengan judul. Selain itu, perlu dilakukan pembersihan data terhadap noise seperti deskripsi yang terlalu singkat, tidak relevan, atau mengandung teks yang tidak bermakna. Kualitas data yang baik akan sangat berpengaruh terhadap hasil analisis yang dihasilkan oleh sistem.

Ketiga, penelitian selanjutnya juga dapat mengembangkan analisis yang tidak hanya berfokus pada kesesuaian teks, tetapi juga pada aspek keaslian dan asal-usul data budaya. Hal ini penting mengingat data budaya memiliki nilai historis dan identitas yang kuat. Sistem dapat dikembangkan untuk mengidentifikasi asal daerah, jenis budaya (misalnya tarian, makanan, atau upacara adat), serta memastikan bahwa informasi yang digunakan sesuai dengan konteks budaya yang sebenarnya. Dengan demikian, sistem tidak hanya mampu menilai kesesuaian teks, tetapi juga berkontribusi dalam menjaga keaslian dan keakuratan informasi budaya.

Keempat, dari sisi *rule-based* validation, penelitian selanjutnya disarankan untuk mengembangkan *rule* yang lebih adaptif dan kontekstual, misalnya dengan menambahkan *rule* yang lebih spesifik terhadap domain budaya Indonesia. Pengembangan *rule* juga dapat diarahkan agar lebih fleksibel dalam menangkap variasi bahasa, termasuk istilah daerah dan perbedaan gaya penulisan.

Kelima, dari sisi implementasi sistem, pengembangan lebih lanjut dapat dilakukan dengan menambahkan fitur-fitur tambahan pada aplikasi web, seperti visualisasi yang lebih interaktif, penyimpanan histori analisis, serta integrasi

dengan data*base* yang lebih luas. Sistem juga dapat dikembangkan menjadi layanan berbasis API agar dapat digunakan pada berbagai platform lainnya.

Terakhir, dari sisi integrasi nilai Islam, penelitian selanjutnya dapat lebih mengembangkan penerapan prinsip kehati-hatian dalam pengolahan informasi, khususnya dalam memastikan bahwa data yang digunakan tidak hanya sesuai secara tekstual, tetapi juga benar dan dapat dipertanggungjawabkan. Hal ini sejalan dengan prinsip *tabayyun* yang menekankan pentingnya ketelitian dalam menerima dan menyampaikan informasi.

DAFTAR PUSTAKA

Amrulloh, A., & Adam, I. F. (2021). Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF-IDF. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, *7*(2), 74. https://doi.org/10.24014/coreit.v7i2.14917

Aubaid, A. M., Mishra, A., & Mishra, A. (2024). Machine learning and rule-based embedding techniques for classifying text documents. *International Journal of System Assurance Engineering and Management*, *15*(12), 5637–5652. https://doi.org/10.1007/s13198-024-02555-w

Ba, Y., Mancenido, M. V., Chiou, E. K., & Pan, R. (2025). *Data Quality in Crowdsourcing and Spamming Behavior Detection* (arXiv:2404.17582). arXiv. https://doi.org/10.48550/arXiv.2404.17582

Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. *IEEE Access*, *9*, 158972–158983. https://doi.org/10.1109/ACCESS.2021.3130902

Chamira, S. (2022). Implementasi Metode Text Mining Frequency-Invers Document Frequency (Tf-Idf) Untuk Monitoring Diskusi Online. *Journal of Informatics, Electrical and Electronics Engineering*, *1*(3), 97–102. https://doi.org/10.47065/jieee.v1i3.353

Elkabalawy, M., Al-Sakkaf, A., Mohammed Abdelkader, E., & Alfalah, G. (2024). CRISP-DM-Based Data-Driven Approach for Building Energy Prediction Utilizing Indoor and Environmental Factors. *Sustainability*, *16*(17), 7249. https://doi.org/10.3390/su16177249

Gagliardi, I., & Artese, M. T. (2024). Exploring and Visualizing Multilingual Cultural Heritage Data Using Multi-Layer Semantic Graphs and Transformers. *Electronics*, *13*(18), 3741. https://doi.org/10.3390/electronics13183741

Julians, A. R., Manongga, D. H. F., & Hendry, H. (2024). Analisis konten budaya kolaboratif berbasis Grounded Theory menggunakan Text Mining. *AITI*, *21*(2), 230–250. https://doi.org/10.24246/aiti.v21i2.230-250

Kaldeli, E., Menis-Mastromichalakis, O., Bekiaris, S., Ralli, M., Tzouvaras, V., & Stamou, G. (2021). CrowdHeritage: Crowdsourcing for Improving the Quality of Cultural Heritage Metadata. *Information*, *12*(2), 64. https://doi.org/10.3390/info12020064

Opitz, J., Möller, L., Michail, A., Padó, S., & Clematide, S. (2025). *Interpretable Text Embeddings and Text Similarity Explanation: A Survey* (arXiv:2502.14862). arXiv. https://doi.org/10.48550/arXiv.2502.14862

Plotnikova, V., Dumas, M., & Milani, F. P. (2022). Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. *Data & Knowledge Engineering*, *139*, 102013. https://doi.org/10.1016/j.datak.2022.102013

Prasad, K. M. S. (2023). Text mining: Identification of similarity of text documents using hybrid similarity model. *Iran Journal of Computer Science*, *6*(2), 123–135. https://doi.org/10.1007/s42044-022-00127-4

Pulungan, B. S. (2024). *Natural Language Processing Ekstraksi Akronim Dan Ekspansi Pada Artikel Berbahasa Indonesia Menggunakan Metode Text Mining Dan Term Frequency-Inverse Document Frequency*. *3*(3).

Putro, M. H., Larasati, D., & Ratri, D. (2022). Eksplorasi Metode Crowdsourcing dalam Upaya Pengarsipan Musik melalui Perancangan Web-Based Director. *ANDHARUPA: Jurnal Desain Komunikasi Visual & Multimedia*, *8*(02), 139–155. https://doi.org/10.33633/andharupa.v8i02.4398

Sari, H., Ginting, G. L., & Zebua, T. (2021). *Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma*. *2*(7).

Sudar, L. Z. S., Imbenay, J. L., Budi, I., Ramadiah, A., Putra, P. K., & Santoso, A. B. (2024). Textual Analysis for Public Sentiment Toward National Police Using CRISP-DM Framework. *Revue d'Intelligence Artificielle*, *38*(1), 63–72. https://doi.org/10.18280/ria.380107

Wibowo, E., & Salim, T. A. (2024). Pengaruh crowdsourcing dan representasi koleksi terhadap kredibilitas informasi pada museum di Indonesia. *Berkala Ilmu Perpustakaan dan Informasi*, *20*(1), 177–192. https://doi.org/10.22146/bip.v20i1.10019

Widianto, A., Pebriyanto, E., Fitriyanti, F., & Marna, M. (2024). Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity. *Journal of Dinda : Data Science, Information Technology, and Data Analytics*, *4*(2), 149–153. https://doi.org/10.20895/dinda.v4i2.1589

Zheng, Y., Li, F., Li, C., Zhang, Z., Cao, R., & Sohail, N. (2024). A Natural Language Processing Model for Automated Organization and Analysis of Intangible Cultural Heritage: *Journal of Organizational and End User Computing*, *36*(1), 1–27. https://doi.org/10.4018/JOEUC.349736