

ORKESTRASI ALUR ANALITIK DATA MULTI-SUMBER TEROTOMASI MENGGUNAKAN APACHE AIRFLOW UNTUK PEMANTAUAN REPUTASI INSTITUSI

Muhamad Fa'iq Awaludin Syiam^{*}, Nono Heryana², Apriade Voutama³

^{1,2,3}Program Studi Sistem Informasi, Fakultas Ilmu Komputer; Universitas Singaperbangsa Karawang, Jl. HS. Ronggowaluyo, Telukjambe Timur, Karawang -41361, Telp/Fax: (0267) 641177

Keywords:

Apache Airflow; Analisis Sentimen; Modern Data Stack; Pemantauan Reputasi; Retrieval-Augmented Generation.

Correspondent Email:

2210631250088@student.unsika.ac.id

Abstrak. Reputasi institusi di era digital dipengaruhi oleh opini publik yang tersebar melalui media berita daring dan media sosial. Pemantauan secara manual terhadap berbagai sumber informasi tersebut memerlukan waktu yang lama dan sulit dilakukan secara berkelanjutan. Penelitian ini bertujuan mengimplementasikan alur analitik data multi-sumber terotomasi menggunakan Apache Airflow untuk mendukung pemantauan reputasi institusi secara lebih efisien. Pengembangan sistem menerapkan metodologi CRISP-DM dan Agile Scrum. Apache Airflow berperan sebagai orkestrator yang mengotomatisasi proses ekstraksi data, transformasi, analisis sentimen menggunakan IndoBERT, penyimpanan data ke PostgreSQL, hingga penyajian hasil melalui dashboard berbasis Streamlit. Sistem juga mengintegrasikan Retrieval-Augmented Generation (RAG) dengan Google Gemini untuk menghasilkan ringkasan analisis dan layanan chatbot berbasis data. Hasil evaluasi menunjukkan bahwa penerapan Hybrid Logic meningkatkan akurasi klasifikasi sentimen dari 50% menjadi 80%, sedangkan evaluasi RAG memperoleh nilai Faithfulness sebesar 1,00 dan Answer Relevancy sebesar 0,98. Sistem yang dikembangkan mampu mengotomatisasi proses pemantauan reputasi serta menyajikan informasi analitik secara cepat sebagai pendukung pengambilan keputusan berbasis data.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract. Institutional reputation in the digital era is increasingly shaped by public opinion disseminated through online news media and social platforms. Manual monitoring of these information sources is time-consuming and difficult to sustain continuously. This study aims to implement an automated multi-source data analytics pipeline using Apache Airflow to support institutional reputation monitoring more efficiently. The system was developed using the CRISP-DM methodology and Agile Scrum framework. Apache Airflow serves as the workflow orchestrator that automates data extraction, transformation, sentiment analysis using IndoBERT, data storage in PostgreSQL, and visualization through a Streamlit dashboard. The system also integrates Retrieval-Augmented Generation (RAG) with Google Gemini to provide AI-generated summaries and a data-driven chatbot. Experimental results show that the proposed Hybrid Logic improves sentiment classification accuracy from 50% to 80%, while the RAG evaluation achieves a Faithfulness score of 1.00 and an Answer Relevancy score of 0.98. The proposed system successfully automates institutional reputation monitoring and delivers analytical insights to support data-driven decision making.

1. PENDAHULUAN

Di era digital saat ini, reputasi institusi perguruan tinggi sangat dipengaruhi oleh dinamika opini publik yang tersebar luas di media berita daring [1]. Selain itu, wacana interaktif di platform media sosial seperti X (Twitter) juga menjadi faktor penentu dalam membentuk persepsi masyarakat [2]. Kumpulan data tekstual berskala masif ini sebenarnya menjadi aset strategis bagi Unit Hubungan Masyarakat (Humas) untuk mendeteksi potensi krisis sejak dini [3]. Sayangnya, operasional pemantauan isu publik di lingkungan institusi umumnya masih dilakukan secara manual dan reaktif [4]. Metode pencatatan manual ini sangat rentan terhadap fenomena *information overload* ketika menghadapi lonjakan data [5]. Kondisi tersebut pada akhirnya memicu inkonsistensi pelabelan opini dan memperlambat langkah mitigasi krisis [6]. Tanpa adanya infrastruktur rekayasa data (*Data Engineering*) yang mampu mengotomatisasi pengumpulan informasi, institusi akan terus mengalami hambatan dalam merespons krisis komunikasi publik secara tepat waktu [7].

Tinjauan terhadap penelitian terdahulu mengenai sistem pemantauan reputasi masih menyisakan celah teknologi yang signifikan. Mayoritas eksperimen saat ini masih berfokus murni pada pengembangan model *Natural Language Processing* (NLP). Model IndoBERT memang terbukti sangat akurat dalam mengklasifikasikan teks bahasa Indonesia secara umum [8]. Namun, penelitian terbaru dari Shulhiyana dkk. mengungkap sebuah kelemahan mendasar pada model ini melalui uji *ablation* [9]. Mereka menemukan bahwa IndoBERT cenderung memprediksi sentimen berdasarkan gaya penulisan makro secara keseluruhan. Akibatnya, berita krisis yang ditulis menggunakan format jurnalistik formal sering kali keliru diklasifikasikan sebagai fakta yang netral [9]. Untuk mengatasi anomali deteksi tersebut, sebuah solusi telah dibuktikan efektivitasnya oleh Manah dkk. [10]. Penelitian mereka menunjukkan bahwa pendekatan hibrida, yaitu dengan menggabungkan kemampuan IndoBERT dan aturan leksikon (*profanity rule*), sangat tangguh dalam mengoreksi bias prediksi sentimen [10]. Meskipun temuan algoritma dari kedua penelitian tersebut sangat inovatif, implementasinya baru mencapai tahap evaluasi

statis dan belum dioperasionalkan ke dalam infrastruktur *DataOps* yang berjalan secara berkesinambungan [11].

Untuk menjembatani kesenjangan tersebut, penelitian ini bertujuan merancang arsitektur *Automated Data Analytics Pipeline* yang beroperasi secara otonom. Sistem ini menggunakan platform Apache Airflow sebagai orkestrator sentral untuk menjalankan proses *Extract, Transform, Load* (ETL) dari berbagai sumber publik tanpa memerlukan intervensi manusia [12]. Kebaruan penelitian ini terletak pada integrasi tiga komponen analitik utama. Pertama, sistem menerapkan klasifikasi sentimen menggunakan IndoBERT berpendekatan hibrida (*rule-based*) terinspirasi dari metode Manah dkk. guna mengoreksi bias model prapelatihan yang diungkap oleh Shulhiyana dkk. Kedua, sistem menyimpan representasi vektor semantik secara *native* langsung pada pangkalan data PostgreSQL menggunakan ekstensi pgvector. Ketiga, arsitektur ini memanfaatkan teknologi *Retrieval-Augmented Generation* (RAG) dari Google Gemini untuk menyajikan ringkasan eksekutif yang akurat dan bebas halusinasi [13]. Melalui integrasi seluruh komponen *Modern Data Stack* tersebut, sistem ini diharapkan mampu menjadi infrastruktur intelijen pendukung keputusan yang tangguh bagi institusi.

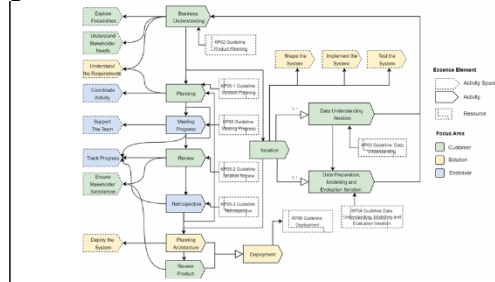
2. TINJAUAN PUSTAKA

2.1. Integrasi Metodologi CRISP-DM dan Agile Scrum

Pengembangan sistem berbasis pembelajaran mesin (*Machine Learning*) memiliki tingkat ketidakpastian yang tinggi. Oleh karena itu, penelitian ini memadukan *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dengan kerangka kerja *Agile Scrum* [14]. Pendekatan ini memastikan bahwa tahapan persiapan data, pemodelan, hingga evaluasi mengikuti standar sains data secara ketat melalui fase CRISP-DM, sementara rilis fungsionalitas sistem dikerjakan secara iteratif dan adaptif melalui siklus *Sprint* pada Scrum [14]. Integrasi kedua metodologi penelitian tersebut divisualisasikan pada

gambar

berikut:



Gambar 1. Integrasi CRISP-DM dan Scrum

2.2. Orkestrasi Pipa Data Terotomasi (Apache Airflow)

Dalam pengelolaan volume data yang masif, keandalan sistem sangat bergantung pada otomatisasi proses *Extract, Transform, Load* (ETL). *Apache Airflow* adalah platform sumber terbuka yang digunakan untuk merancang, menjadwalkan, dan memantau alur kerja komputasi yang direpresentasikan sebagai *Directed Acyclic Graphs* (DAGs). Dalam arsitektur *Modern Data Stack*, implementasi Airflow bertindak sebagai orkestrator sentral yang memastikan penarikan data dari antarmuka pemrograman aplikasi (API) eksternal berjalan secara paralel dan berkesinambungan tanpa intervensi manual [12].

2.3. Pemodelan IndoBERT dan Logika Hybrid

IndoBERT merupakan model bahasa berbasis arsitektur *Transformer* yang dilatih menggunakan korpus teks bahasa Indonesia berskala masif, sehingga unggul dalam representasi pemahaman konteks dua arah [15]. Namun, model prapelatihan murni rentan terhadap bias gaya penulisan berita jurnalistik sehingga kerap keliru dalam mendeteksi konteks krisis spesifik [9]. Sebagai solusinya, diterapkan logika klasifikasi hibrida yang mengombinasikan ketajaman semantik *Transformer* dengan aturan leksikon (*rule-based*), sebuah pendekatan yang terbukti secara komputasional mampu mengoreksi bias prediksi [10].

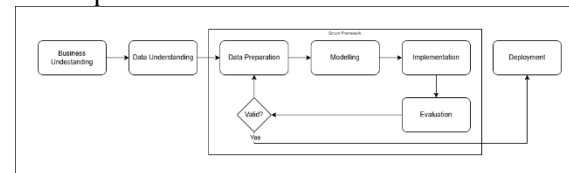
2.4. Native Vector Database (pgvector) dan RAG

Kecerdasan buatan generatif (*Generative AI*) rentan terhadap fenomena halusinasi data apabila tidak dibatasi oleh fakta empiris.

Kerangka kerja *Retrieval-Augmented Generation* (RAG) mengatasi celah ini dengan cara menyuplai konteks dokumen lokal yang relevan kepada *Large Language Model* (LLM) [16]. Untuk mendukung pencarian kesamaan jarak spasial (*semantic search*) dengan latensi rendah, representasi vektor matematis (*embeddings*) disimpan secara *native* berdampingan dengan data relasional pada pangkalan data PostgreSQL menggunakan ekstensi *pgvector* [13].

3. METODE PENELITIAN

Metode penelitian ini mengadaptasi kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang diintegrasikan dengan metode *Agile Scrum* guna mengakomodasi pembangunan sistem *Machine Learning Operations* (MLOps) secara iteratif. Tahapan penelitian ini difokuskan pada pemrosesan rekayasa data (*Data Engineering*) yang mencakup fase persiapan data (*Data Preparation*), pemodelan (*Modeling*), hingga penyebaran sistem (*Deployment*). Alur tahapan metodologi penelitian tersebut disajikan secara visual pada Gambar 2.



Gambar 2. Rancangan Penelitian

3.1. Pengumpulan Data dan Orkestrasi

Proses pengumpulan data menyasar wacana publik terkait institusi dari portal berita (Google News) dan platform media sosial X [1]. Pengumpulan data dilakukan menggunakan teknik *web scraping* otomatis. Ekstraksi portal berita dibangun menggunakan pustaka Python BeautifulSoup dengan filter waktu *when:1y*, sedangkan ekstraksi media sosial X menggunakan pustaka *tweet-harvest*.

Seluruh proses ekstraksi diorkestrasi menggunakan *Apache Airflow* melalui *Directed Acyclic Graph* (DAG) *remosy_master_end_to_end*. *Scheduler* diatur untuk mengeksekusi penarikan data secara paralel setiap pukul 00.00 WIB dan 12.00 WIB. Sebelum data mentah disimpan ke dalam lapisan *Bronze* (tabel *news_raw* dan *tweets_raw* pada PostgreSQL), sistem menjalankan

algoritma *Check-Sum URL* untuk menolak data duplikat.

3.2. Pra-pemrosesan Data

Data mentah yang terkumpul selanjutnya memasuki tahap pembersihan teks. Proses pra-pemrosesan dieksekusi menggunakan pustaka *Regular Expression* (re) pada kerangka kerja Pandas. Tahapan pembersihan meliputi:

- 1) Penghapusan tautan hiperteks (URL) dan *mention* akun (@)
- 2) Eliminasi karakter non-alfabetik, tanda baca berlebih, dan *emoji*
- 3) Penyeragaman format huruf menjadi huruf kecil seluruhnya (*case folding*)

3.3. Pemodelan Sentimen

Teks yang telah dibersihkan kemudian diklasifikasikan ke dalam tiga kelas sentimen (Positif, Negatif, Netral). Pemodelan dilakukan menggunakan model prapelatihan indonesian-roberta-base-sentiment-classifier dari *HuggingFace*. Teks masukan dibatasi maksimal 512 token untuk menghindari luapan memori (*memory overflow*).

Hasil prediksi dasar dari model prapelatihan kemudian dikalibrasi menggunakan aturan heuristik (*Rule-Based Logic*). Algoritma ini memindai kemunculan diksi spesifik yang mengindikasikan krisis institusi (seperti: "korupsi", "pungli", atau "demo"). Jika terdeteksi, sistem akan secara otomatis menimpa (*override*) prediksi awal model menjadi sentimen Negatif. Hasil akhir dari proses ini disimpan ke dalam tabel `news_processed` dan `tweets_processed`.

3.4. Vektorisasi dan Retrieval-Augmented Generation

Pada lapisan *Gold*, teks berlabel dikonversi menjadi representasi vektor numerik (*embeddings*). Vektor tersebut disimpan secara *native* berdampingan dengan data relasional pada PostgreSQL menggunakan ekstensi *pgvector*.

Untuk menyajikan asisten cerdas pada dasbor antarmuka Streamlit, sistem mengimplementasikan *Retrieval-Augmented Generation* (RAG). Saat sistem menerima kueri dari pengguna, pencarian konteks dokumen terdekat dilakukan dengan menghitung jarak kemiripan (*Cosine Similarity*) antara vektor

kueri (Q) dan vektor dokumen (D) menggunakan persamaan:

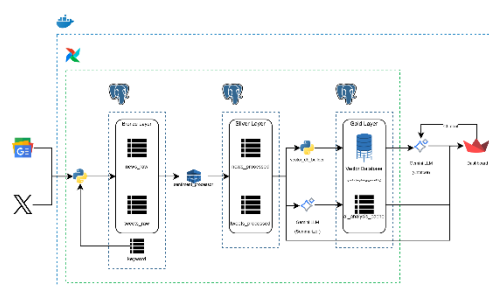
$$\text{sim}(Q, D) = \frac{Q \cdot D}{\|Q\| \|D\|} \quad (1)$$

Dokumen historis yang memiliki skor kemiripan tertinggi akan diekstraksi dan disuntikkan sebagai konteks absolut ke dalam instruksi *prompt* pada *Large Language Model* (Google Gemini 2.5-Flash) untuk memproduksi ringkasan eksekutif dan jawaban obrolan yang faktual.

4. HASIL DAN PEMBAHASAN

4.1. Orkestrasi Workflow Menggunakan Apache Airflow

Apache Airflow berperan sebagai pusat orkestrasi (*workflow orchestrator*) yang mengendalikan seluruh proses pada *pipeline*, mulai dari akuisisi data, transformasi, analisis sentimen, pembentukan *embedding*, hingga pembaruan basis data dan dashboard secara otomatis. Seluruh proses dirancang agar mampu mengumpulkan data dari berbagai sumber secara terjadwal, melakukan transformasi data, menjalankan analisis berbasis kecerdasan buatan, serta menyajikan hasil analitik kepada pengguna melalui *dashboard* dan *chatbot*. Pendekatan ini memungkinkan proses pemantauan reputasi institusi dilakukan secara berkelanjutan tanpa intervensi manual.

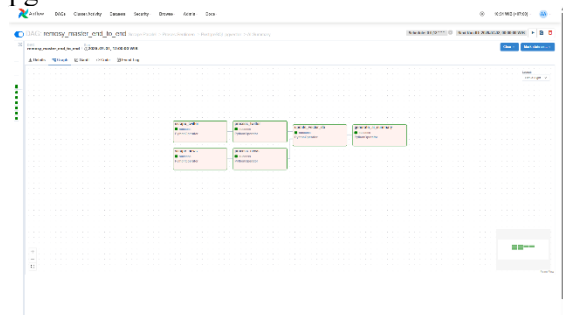


Gambar 3. Arsitektur Sistem

Berdasarkan Gambar 3, sistem dibangun menggunakan lima lapisan utama, yaitu *Data Source*, *Data Engineering Layer*, *Storage Layer*, *AI Analytics Layer*, dan *Presentation Layer*. Data dikumpulkan dari Google News dan platform X, kemudian diproses melalui *pipeline* ETL menggunakan Apache Airflow. Data hasil transformasi disimpan pada PostgreSQL sebagai basis data utama, sedangkan ekstensi *pgvector* digunakan untuk

menyimpan *embedding vector* yang mendukung mekanisme *Retrieval-Augmented Generation* (RAG). Pada lapisan analitik, model IndoBERT digunakan untuk klasifikasi sentimen, sementara Google Gemini memanfaatkan konteks yang diperoleh melalui RAG untuk menghasilkan ringkasan dan menjawab pertanyaan pengguna. Seluruh hasil analisis kemudian disajikan melalui aplikasi berbasis Streamlit sehingga pengguna dapat memantau kondisi reputasi institusi secara interaktif.

Implementasi arsitektur tersebut direalisasikan menggunakan Apache Airflow sebagai *workflow orchestrator* yang memodelkan seluruh proses ke dalam bentuk *Directed Acyclic Graph* (DAG). Setiap *task* merepresentasikan tahapan ETL yang saling bergantung sehingga proses ekstraksi, transformasi, analisis, hingga pembaruan *knowledge base* dapat dijalankan secara otomatis sesuai urutan eksekusinya. Untuk menjaga kualitas data, *pipeline* menerapkan mekanisme *URL Check-Sum* guna menghindari data duplikat, dilanjutkan dengan proses *text pre-processing*, klasifikasi sentimen, pembentukan *embedding vector*, dan penyimpanan hasil ke PostgreSQL serta pgvector.



Gambar 4. *Directed Acyclic Graph*

Gambar 4 menunjukkan implementasi DAG yang digunakan untuk mengorkestrasi seluruh proses ETL. Proses akuisisi data dari Google News dan platform X dijalankan secara paralel melalui dua *task* terpisah sehingga mampu mengurangi waktu ekstraksi dibandingkan pendekatan *sequential execution*. Setelah kedua proses selesai, Airflow melakukan sinkronisasi melalui *join task* sebelum melanjutkan ke tahapan transformasi dan analisis. Seluruh *task* kemudian dijalankan secara terjadwal setiap 12 jam menggunakan *scheduler* Apache Airflow, sehingga sistem mampu memperbarui data,

knowledge base, serta hasil analisis secara otomatis tanpa memerlukan intervensi pengguna.

Implementasi tersebut menunjukkan bahwa Apache Airflow mampu mengorkestrasi seluruh alur analitik data secara terintegrasi, mulai dari pengumpulan data multi-sumber hingga penyediaan informasi yang siap digunakan oleh dashboard dan chatbot. Integrasi PostgreSQL dengan pgvector juga menyederhanakan arsitektur penyimpanan karena data relasional dan *embedding vector* dapat dikelola dalam satu platform basis data, sehingga mendukung efisiensi implementasi sekaligus mempermudah pengembangan sistem pada penelitian ini. Dengan demikian, Apache Airflow tidak hanya berfungsi sebagai *scheduler*, tetapi juga sebagai pusat orkestrasi yang mengintegrasikan seluruh komponen pada *pipeline* analitik data, mulai dari akuisisi data hingga penyajian hasil kepada pengguna.

4.2. Evaluasi Model Analisis Sentimen

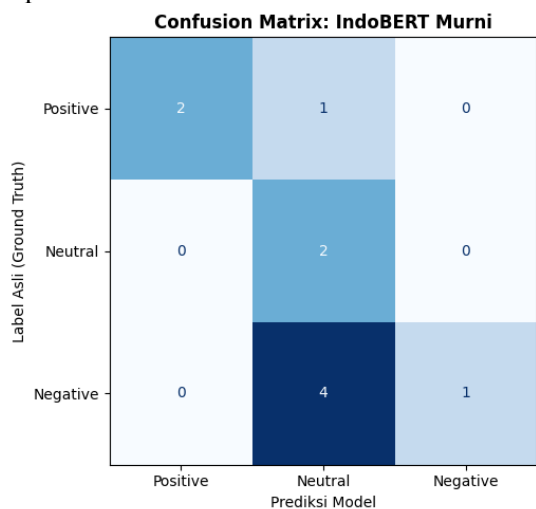
Evaluasi model dilakukan untuk mengukur kinerja klasifikasi sentimen terhadap data berita dan media sosial yang telah melalui tahap *pre-processing*. Penelitian ini membandingkan dua pendekatan, yaitu model IndoBERT sebagai model dasar (*baseline*) dan pendekatan Hybrid Logic, yaitu integrasi antara prediksi IndoBERT dengan aturan berbasis kata kunci (*rule-based*). Kinerja kedua model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* yang diperoleh dari *Confusion Matrix*.

Tabel 1. Perbandingan Kinerja Model Analisis Sentimen

Metrik	IndoBERT	Hybrid Logic
Accuracy	0,50	0,80
Precision	0,86	0,90
Recall	0,50	0,80
F1-Score	0,50	0,82

Berdasarkan Tabel 1, pendekatan *Hybrid Logic* menghasilkan peningkatan performa pada seluruh metrik evaluasi dibandingkan model IndoBERT murni. Akurasi meningkat dari 50% menjadi 80%, sedangkan nilai *recall* meningkat dari 50% menjadi 80%. Peningkatan tersebut menunjukkan bahwa mekanisme aturan (*rule-based*) mampu membantu model mengenali lebih banyak data yang memiliki sentimen sesuai dengan kondisi sebenarnya,

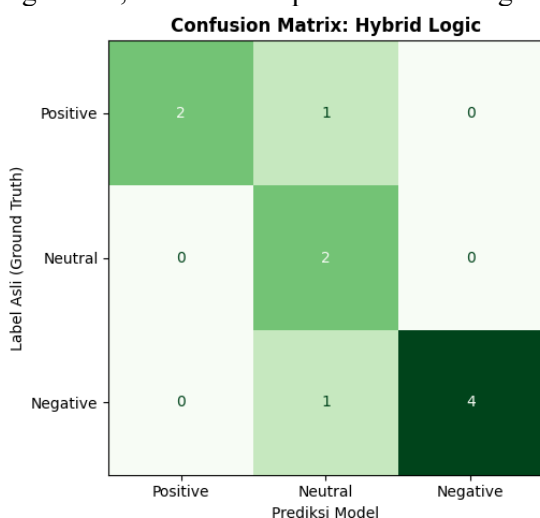
khususnya pada isu yang berkaitan dengan reputasi institusi.



Gambar 5. *Confussion Matrix* IndoBERT

Berdasarkan Gambar 5, model IndoBERT masih mengalami kesalahan klasifikasi pada kelas negatif, yang ditunjukkan oleh rendahnya jumlah prediksi benar (*True Positive*) dan tingginya jumlah data yang salah diklasifikasikan sebagai sentimen netral (*False Negative*). Kondisi ini menyebabkan nilai *recall* pada kelas negatif hanya mencapai 20%, sehingga masih terdapat potensi isu reputasi yang tidak berhasil terdeteksi oleh sistem.

Setelah diterapkan pendekatan Hybrid Logic, distribusi prediksi pada *Confussion Matrix* menunjukkan peningkatan yang signifikan, terutama pada kelas negatif.



Gambar 6. *Confussion Matrix* Hybrid Logic

Integrasi aturan berbasis kata kunci membantu sistem memperbaiki kesalahan

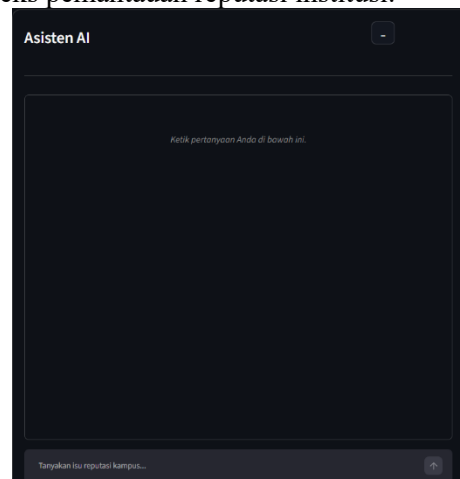
klasifikasi yang sebelumnya dilakukan oleh model IndoBERT sehingga jumlah *False Negative* berkurang.

Dampaknya, nilai *recall* untuk kelas negatif meningkat dari 20% menjadi 80%, sementara akurasi keseluruhan sistem meningkat menjadi 80%.

Hasil evaluasi tersebut menunjukkan bahwa pendekatan hibrida lebih sesuai diterapkan pada sistem pemantauan reputasi institusi dibandingkan penggunaan model *pre-trained* secara mandiri. Dalam konteks pemantauan reputasi, kemampuan mendeteksi isu negatif sejak dini memiliki peran yang lebih penting daripada sekadar memperoleh akurasi tinggi. Oleh karena itu, integrasi antara model *Deep Learning* dan aturan berbasis domain mampu meningkatkan sensitivitas sistem terhadap isu yang berpotensi berkembang menjadi krisis komunikasi.

4.3. Implementasi dan Evaluasi Chatbot Berbasis Retrieval-Augmented Generation

Untuk meningkatkan kemudahan akses terhadap hasil analisis, penelitian ini mengimplementasikan chatbot berbasis *Retrieval-Augmented Generation* (RAG). Berbeda dengan chatbot konvensional yang hanya mengandalkan pengetahuan bawaan *Large Language Model*, pendekatan RAG mengombinasikan kemampuan Google Gemini dengan proses *retrieval* terhadap dokumen yang tersimpan pada PostgreSQL menggunakan ekstensi pgvector. Dengan demikian, jawaban yang dihasilkan mengacu pada data hasil analisis sentimen yang tersedia pada sistem sehingga lebih faktual dan sesuai dengan konteks pemantauan reputasi institusi.



Gambar 7. *Interface Chatbot*

Berdasarkan Gambar 7, pengguna dapat mengajukan pertanyaan mengenai tren sentimen, isu yang sedang berkembang, maupun ringkasan kondisi reputasi institusi melalui bahasa alami. Setiap pertanyaan terlebih dahulu dikonversi menjadi *embedding vector* untuk mencari dokumen yang memiliki kemiripan semantik tertinggi pada basis pengetahuan. Dokumen yang berhasil ditemukan kemudian digunakan sebagai konteks tambahan (*retrieval context*) sebelum diproses oleh Google Gemini sehingga jawaban yang dihasilkan tetap mengacu pada informasi yang tersedia dalam sistem.

Kinerja mekanisme RAG dievaluasi menggunakan metrik *Mean Reciprocal Rank (MRR)*, *Hit Rate*, *Precision@5*, *Faithfulness*, dan *Answer Relevancy*. Metrik tersebut digunakan untuk mengukur kualitas proses *retrieval* maupun kualitas jawaban yang dihasilkan oleh *Large Language Model*.

Tabel 2. Hasil Evaluasi RAG

Dimensi	Metrik	Nilai
Retrieval	Mean Reciprocal Rank (MRR)	0,44
	Hit Rate	0,67
	Precision@5	0,13
Generation	Faithfulness	1,00
	Answer Relevancy	0,98

Berdasarkan Tabel 2, mekanisme *retrieval* mampu menemukan dokumen yang relevan dengan nilai *Hit Rate* sebesar 0,67, sedangkan nilai *MRR* sebesar 0,44 menunjukkan bahwa dokumen yang relevan umumnya muncul pada urutan awal hasil pencarian. Meskipun nilai *Precision@5* masih relatif rendah, hasil tersebut menunjukkan bahwa tidak seluruh dokumen pada lima hasil teratas memiliki tingkat relevansi yang sama sehingga masih terdapat peluang untuk mengoptimalkan proses pencarian dokumen.

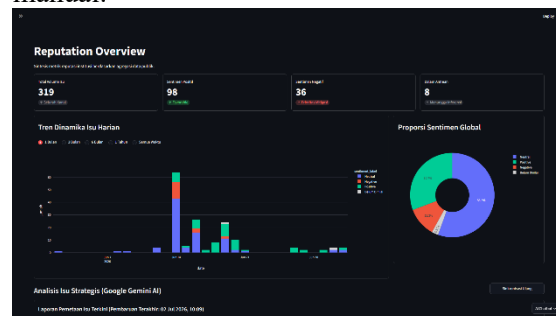
Pada tahap *generation*, sistem memperoleh nilai *Faithfulness* sebesar 1,00 dan *Answer Relevancy* sebesar 0,98. Hasil tersebut menunjukkan bahwa jawaban yang dihasilkan Google Gemini tetap konsisten dengan konteks dokumen yang diperoleh melalui mekanisme *retrieval* serta memiliki tingkat kesesuaian yang sangat tinggi terhadap pertanyaan pengguna.

Dengan demikian, implementasi RAG berhasil mengurangi potensi *hallucination* karena model menghasilkan respons berdasarkan informasi yang tersedia pada basis pengetahuan, bukan semata-mata dari pengetahuan parametriknya.

Secara keseluruhan, integrasi RAG memberikan nilai tambah terhadap sistem pemantauan reputasi yang dikembangkan. Selain menyajikan visualisasi melalui dashboard, sistem juga mampu menyediakan layanan tanya jawab berbasis data yang interaktif, faktual, dan relevan sehingga dapat membantu pengguna memperoleh informasi secara lebih cepat untuk mendukung proses pengambilan keputusan.

4.4. Implementasi Dashboard Monitoring Berbasis Streamlit

Dashboard monitoring dikembangkan menggunakan Streamlit sebagai *presentation layer* untuk menyajikan hasil analisis sentimen dalam bentuk visualisasi interaktif dan informasi tekstual yang mudah dipahami. Dashboard terhubung langsung dengan basis data PostgreSQL sehingga setiap hasil pemrosesan *pipeline* dapat ditampilkan secara otomatis tanpa memerlukan pembaruan manual.

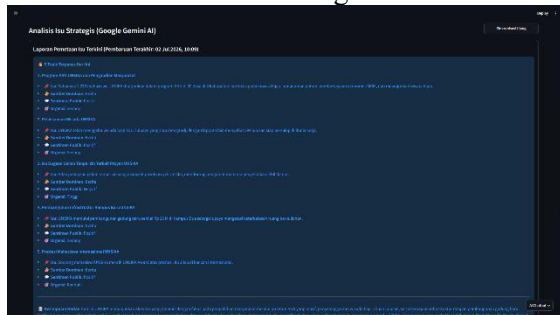


Gambar 8. Dashboard Utama

Berdasarkan Gambar 8, *dashboard* menyajikan indikator utama reputasi institusi yang meliputi total volume isu, distribusi sentimen positif dan negatif, tren dinamika isu harian, serta proporsi sentimen global. Pengguna juga dapat melakukan penyaringan berdasarkan rentang waktu sehingga perubahan sentimen dapat diamati secara lebih fleksibel. Seluruh visualisasi diperbarui secara otomatis mengikuti hasil eksekusi *pipeline* Apache Airflow.

Selain menyajikan visualisasi statistik, dashboard mengintegrasikan layanan Google Gemini untuk menghasilkan ringkasan kondisi

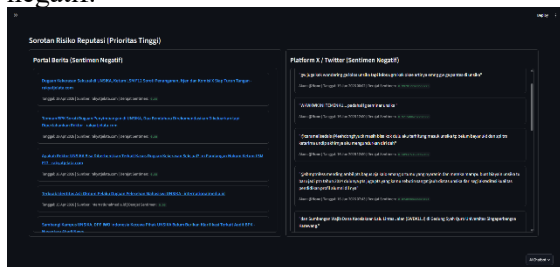
reputasi secara otomatis. Berbeda dengan pendekatan *text summarization* konvensional, proses ini tidak merangkum seluruh dokumen berita secara langsung, tetapi terlebih dahulu mengekstraksi hasil analisis yang telah diproses sistem, seperti kata kunci utama, distribusi sentimen, topik dominan, tingkat urgensi, dan sumber data. Informasi tersebut kemudian digunakan sebagai konteks (*prompt context*) bagi Large Language Model untuk menghasilkan laporan analisis yang lebih terstruktur dan relevan dengan kondisi aktual.



Gambar 9. Ringkasan AI

Berdasarkan Gambar 9, sistem mampu menghasilkan ringkasan otomatis yang memuat lima isu utama, sumber dominan, kecenderungan sentimen publik, tingkat urgensi, serta kesimpulan umum mengenai kondisi reputasi institusi. Pendekatan ini menghasilkan informasi yang lebih mudah dipahami oleh pengguna dibandingkan hanya menampilkan kumpulan data mentah, sehingga dapat membantu proses pemantauan dan pengambilan keputusan.

Sebagai mekanisme pendukung mitigasi risiko, dashboard juga menyediakan halaman khusus yang menampilkan isu dengan prioritas tinggi berdasarkan hasil analisis sentimen negatif.



Gambar 10. Panel Sorotan Negatif

Pada Gambar 10, sistem secara otomatis mengelompokkan berita dan unggahan media sosial yang memiliki kecenderungan sentimen negatif tinggi berdasarkan kata kunci yang

dipantau. Informasi yang ditampilkan meliputi judul berita, isi unggahan, tanggal publikasi, sumber data, serta skor sentimen sehingga pengguna dapat mengidentifikasi isu yang berpotensi memengaruhi reputasi institusi secara lebih cepat. Penyajian ini memungkinkan unit humas memfokuskan perhatian pada isu-isu yang memerlukan tindak lanjut tanpa harus menelusuri seluruh data secara manual.

4.5. Pembahasan

Hasil penelitian menunjukkan bahwa integrasi Apache Airflow, PostgreSQL, IndoBERT, Google Gemini, dan Streamlit mampu membentuk sebuah *automated data analytics pipeline* yang mendukung proses pemantauan reputasi institusi secara terintegrasi. Seluruh tahapan, mulai dari akuisisi data, analisis sentimen, hingga penyajian informasi, dapat berjalan secara otomatis sehingga mengurangi ketergantungan terhadap proses pemantauan manual.

Penerapan *Hybrid Logic* berhasil meningkatkan kemampuan model dalam mendeteksi sentimen negatif, sedangkan implementasi *Retrieval-Augmented Generation* (RAG) memungkinkan Google Gemini menghasilkan ringkasan dan jawaban yang tetap mengacu pada data hasil pemantauan. Integrasi kedua pendekatan tersebut menghasilkan informasi yang lebih akurat dan mudah dipahami oleh pengguna.

Secara keseluruhan, sistem yang dikembangkan tidak hanya berfungsi sebagai media visualisasi hasil analisis, tetapi juga sebagai alat bantu pengambilan keputusan melalui penyajian indikator reputasi, ringkasan isu, dan chatbot berbasis kecerdasan buatan. Hal ini menunjukkan bahwa pendekatan *Modern Data Stack* dapat menjadi alternatif yang efektif dalam membangun sistem pemantauan reputasi institusi berbasis data.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat diambil beberapa kesimpulan sebagai berikut.

- Penelitian ini berhasil mengimplementasikan *automated data analytics pipeline* berbasis Apache Airflow yang mengintegrasikan proses akuisisi data, penyimpanan pada PostgreSQL, analisis

sentimen menggunakan IndoBERT, serta penyajian hasil melalui dashboard Streamlit. Arsitektur yang dibangun mampu mengotomatisasi proses pemantauan reputasi institusi dari berbagai sumber data secara terintegrasi.

- b. Integrasi *Hybrid Logic* pada analisis sentimen serta penerapan *Retrieval-Augmented Generation* (RAG) dengan Google Gemini berhasil meningkatkan kualitas informasi yang dihasilkan sistem. Dashboard tidak hanya menyajikan visualisasi sentimen, tetapi juga menyediakan ringkasan isu strategis dan layanan chatbot berbasis data sehingga dapat mendukung proses pemantauan reputasi dan pengambilan keputusan secara lebih efektif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada kedua orang tua atas doa, kasih sayang, dan dukungan yang tiada henti selama proses penyusunan penelitian ini. Penulis juga menyampaikan penghargaan yang sebesar-besarnya kepada dosen pembimbing atas bimbingan, arahan, serta masukan yang sangat berarti dalam setiap tahapan penelitian. Ucapan terima kasih turut disampaikan kepada seluruh pihak yang telah membantu dan memberikan dukungan sehingga penelitian dan penulisan artikel ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

- [1] R. Trisnawati, K. Kompri, D. Isnanto, and R. Wenny Asriani, "Enhancing Brand Awareness in Private Islamic Universities: The Peso Model in Jambi Province," *J. Educ. J. Educ. Stud.*, vol. 9, no. 1, pp. 56–69, Aug. 2024, doi: 10.30983/educative.v9i1.8502.
- [2] F. A. Prakosa and D. E. Permanasari, "Penggunaan Instagram @umkla.official sebagai Sarana Branding dan Promosi Kampus," vol. 2, no. 1, 2024.
- [3] R. Andi Kambau, "Proses Transformasi Digital pada Perguruan Tinggi di Indonesia," *J. Rekayasa Sist. Inf. Dan Teknol.*, vol. 1, no. 3, pp. 126–136, Feb. 2024, doi: 10.59407/jrsit.v1i3.481.
- [4] I. R. Wibowo *et al.*, "BIG DATA PIPELINE INFRASTRUCTURE DESIGN IN MSME E-COMMERCE SYSTEMS WITH A FOCUS ON DATA SOURCE PROCESSING USING ORCHESTRATION TOOLS," *Transm. J. Ilm. Tek. Elektro*, vol. 26, no. 1, pp. 48–54, Jan. 2024, doi: 10.14710/transmisi.26.1.48-54.
- [5] R. H. Astuti, M. Muljono, and S. Sutriawan, "Indonesian News Text Summarization Using MBART Algorithm," *Sci. J. Inform.*, vol. 11, no. 1, pp. 155–164, Feb. 2024, doi: 10.15294/sji.v11i1.49224.
- [6] I. C. Tresna and R. Sijabat, "Analisis Pengaruh Digital Marketing, Brand Awareness dan Campus Facilities terhadap Enrollment Intention pada Sebuah Perguruan Tinggi Swasta," *WIDYAKALA J. J. Pembang. JAYA Univ.*, vol. 10, no. 2, p. 87, Oct. 2023, doi: 10.36262/widyakala.v10i2.831.
- [7] A. K. Dipa, H. Hafiar, and A. Rahmat, "Pemahaman Humas Lembaga Perguruan Tinggi Tentang Reputasi di Era Digital Public Relations," *J. Audience*, vol. 5, no. 2, pp. 209–225, Feb. 2023, doi: 10.33633/ja.v5i2.6190.
- [8] M. I. Abidin and E. W. Pamungkas, "ANALISIS SENTIMEN TERHADAP TIMNAS INDONESIA DI PIALA ASIA 2023 DENGAN MODEL TRANSFORMER BERBAHASA INDONESIA," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 482–496, Jul. 2025, doi: 10.36341/rabit.v10i2.6142.
- [9] S. Shulhiyana, N. Heryana, and T. Ridwan, "PERBANDINGAN KINERJA MODEL BILSTM DAN INDOBERT DALAM MENDETEKSI BERITA HOAKS PADA MEDIA ONLINE DI INDONESIA".
- [10] A. S. Manah, "Analisis Sentimen Publik Terhadap Kasus Keracunan Makanan MBG Berbasis Data Media Sosial Menggunakan Model Indobert," *J. Inform. Dan Tek. Elektro Terap.*, vol. 14, no. 2, Apr. 2026, doi: 10.23960/jitet.v14i2.9404.
- [11] F. Dea Reskyadita, H. Nurrahmi, and D. Maulana, "Design and Implementation of a Performance Dashboard for Public Relations at Telkom University: Perancangan dan Implementasi Dashboard Kinerja Bidang Humas Universitas Telkom," *NUANSA Inform.*, vol. 19, no. 2, pp. 133–143, Jul. 2025, doi: 10.25134/ilkom.v19i2.377.
- [12] E. E. Wahyudi *et al.*, "Akuisisi Data Prediksi Curah Hujan Secara Periodik Menggunakan Apache Airflow," *J. Inform. Inf. Syst. Softw. Eng. Appl. INISTA*, vol. 4, no. 2, pp. 1–12, May 2022, doi: 10.20895/inista.v4i2.574.
- [13] A. N. Hadie, I. Tahyudin, and T. Hariguna, "Enhancing Accessibility in Local Government Data Portals via Retrieval-Augmented Generation: A Case Study on Satu Data Indonesia in Banyumas Regency," *J. Tek. Inform. Jutif*, vol. 6, no. 4, pp. 2420–2433, Sep. 2025, doi: 10.52436/1.jutif.2025.6.4.5153.
- [14] M. T. Hayat Suhendar and Y. Widayani, "Machine Learning Application Development Guidelines Using CRISP-DM and Scrum

- Concept,” in *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)*, Toba, Indonesia: IEEE, Sep. 2023, pp. 168–173. doi: 10.1109/ICoDSE59534.2023.10291438.
- [15] A. A. Qolbu, N. Fitriyati, and N. Inayah, “Performa Naïve Bayes, SVM, dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang,” *J. FOURIER*, 2025.
- [16] Rafli Pasya Author, Yuyun Umaidah Author, and Mohamad Jajuli Author, “RANCANG BANGUN SISTEM TANYA JAWAB AKADEMIK DOSEN DENGAN METODE RETRIEVAL AUGMENTED GENERATION BERBASIS WEBSITE,” *J. Inform. Dan Tek. Elektro Terap.*, vol. 13, no. 3S1, Oct. 2025, doi: 10.23960/jitet.v13i3S1.7614.