

PENGEMBANGAN CHATBOT INAPINEP.ID BERBASIS NATURAL LANGUAGE PROCESSING SEBAGAI SISTEM REKOMENDASI HOTEL DI PULAU JAWA

Siti Naia Hesti Rachmawati

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
e-mail: 24083010047@student.upnjatim.ac.id*

Fanny Widya Cahyani

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
e-mail: 24083010045@student.upnjatim.ac.id

Sonya Audina Akbar

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
e-mail: 24083010013@student.upnjatim.ac.id

Shindi Shella May Wara

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
e-mail: shindi.shella.fasikom@upnjatim.ac.id

Amri Muhaimin

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur
e-mail: amri.muhaimin.stat@upnjatim.ac.id

Abstrak

Pesatnya pertumbuhan sektor pariwisata di Pulau Jawa menyebabkan terjadinya fenomena *information overload* bagi wisatawan dalam menentukan akomodasi yang sesuai. Penelitian ini bertujuan untuk mengembangkan asisten virtual bernama *inapinep.id*, sebuah *chatbot* berbasis *Natural Language Processing* (NLP) yang berfungsi sebagai sistem rekomendasi hotel. Metode yang digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk pembobotan fitur dan *Cosine Similarity* untuk menghitung tingkat kemiripan antara kueri deskriptif pengguna dengan dataset hotel. Tahapan *preprocessing* difokuskan pada pembersihan teks, normalisasi sinonim, dan deduplikasi kata tanpa menggunakan proses *stemming* guna menjaga integritas istilah teknis fasilitas hotel. Hasil penelitian menunjukkan bahwa sistem mampu memberikan rekomendasi yang presisi dengan skor kemiripan tertinggi mencapai 0,6687 pada pengujian kueri spesifik. Uji statistik *Chi-Square* mengonfirmasi bahwa fitur "lift" merupakan parameter pembeda paling signifikan ($p = 0,0071$) dalam menentukan kecenderungan rating hotel tinggi dalam dataset. Sistem juga dilengkapi dengan mekanisme *guardrails* untuk menangani kueri di luar topik guna menjaga kualitas interaksi. Penelitian ini membuktikan bahwa integrasi model pembobotan kata dan analisis statistik mampu menghasilkan sistem rekomendasi yang cerdas dan intuitif dalam membantu navigasi wisatawan di Pulau Jawa.

Kata Kunci: *Chatbot, Natural Language Processing, TF-IDF, Cosine Similarity, Rekomendasi Hotel.*

Abstract

The rapid growth of the tourism sector on the island of Java has led to a phenomenon of *information overload* for tourists when selecting suitable accommodations. This study aims to develop a virtual assistant named *inapinep.id*, a *Natural Language Processing* (NLP)-based *chatbot* that functions as a hotel recommendation system. The methods used include *Term Frequency-Inverse Document Frequency* (TF-IDF) for feature weighting and *Cosine Similarity* to calculate the similarity level between users' descriptive queries and the hotel dataset. The preprocessing stage focused on text cleaning, synonym normalization, and word deduplication without using *stemming* to preserve the integrity of technical hotel facility terms. The research results show that the system is capable of providing precise recommendations, with the highest similarity score reaching 0,6687 in specific query testing. A *Chi-Square* statistical test confirms that the "lift" feature is the most significant distinguishing parameter ($p = 0,0071$) in determining the tendency for high hotel ratings in the dataset. The system is also equipped with *guardrail* mechanisms to handle off-topic queries to maintain interaction quality. This research demonstrates that the integration of word weighting models and statistical analysis can produce an intelligent and intuitive recommendation system to assist travelers navigating Java.

Keywords: Chatbot, Natural Language Processing, TF-IDF, Cosine Similarity, Hotel Recommendation.

PENDAHULUAN

Kemajuan sektor pariwisata di Indonesia menjadi salah satu pelopor utama pertumbuhan ekonomi nasional. Pariwisata berkontribusi terhadap peningkatan pendapatan daerah serta membuka peluang usaha di bidang akomodasi seperti hotel dan motel yang semakin beragam. Pengembangan sektor pariwisata didukung infrastruktur dan teknologi digital yang dinilai mampu meningkatkan daya tarik destinasi serta mempermudah akses informasi bagi wisatawan (Firanti, 2026).

Dari perspektif yang berbeda, banyaknya pilihan hotel dan motel justru memicu permasalahan baru bagi masyarakat. Pengguna sering mengalami kesulitan dalam menentukan akomodasi yang sesuai dengan lokasi, anggaran, dan fasilitas yang diinginkan (Hananto, 2023). Informasi hotel umumnya tersebar di berbagai platform digital maupun media sosial sehingga proses pencarian menjadi kurang efektif dan membutuhkan waktu yang lama. Pemanfaatan teknologi digital dalam sektor pariwisata menjadi solusi yang terus berkembang, salah satunya melalui sistem rekomendasi berbasis kecerdasan buatan. Penelitian oleh Wulandari et al. (2025) menunjukkan bahwa aspek aksesibilitas informasi dan fasilitas pendukung merupakan faktor penting dalam pengembangan destinasi wisata.

Salah satu pendekatan teknologi yang dapat digunakan adalah chatbot berbasis *Natural Language Processing* (NLP). Chatbot merupakan implementasi kecerdasan buatan yang mensimulasikan percakapan secara *real-time* antara mesin dan manusia melalui teks, sehingga memungkinkan pengguna berinteraksi menggunakan bahasa alami dan proses pencarian informasi menjadi lebih tanggap (Chandra et al., 2022). Penerapan NLP pada chatbot telah terbukti efektif di berbagai bidang layanan, termasuk layanan informasi akademik dengan tingkat akurasi rata-rata mencapai 91%, yang menunjukkan potensi besar teknologi ini untuk diadaptasi pada konteks layanan informasi lainnya (Chandra et al., 2022). Penelitian oleh Simorangkir et al. (2024) menunjukkan bahwa chatbot rekomendasi hotel berbasis NLP mampu membantu pengguna

memperoleh informasi hotel secara lebih efisien melalui percakapan otomatis. Selain itu, metode *Named Entity Recognition* (NER) dan *Cosine Similarity* juga terbukti efektif dalam memahami perihwal pertanyaan pengguna serta mencocokkan kebutuhan pengguna dengan data rekomendasi hotel yang tersedia dalam dataset (Kusumaningtyas et al., 2022).

Pengembangan chatbot berbasis NLP semakin diperkuat dengan hadirnya teknik NER yang memungkinkan sistem mengenali entitas penting seperti lokasi, nama layanan, dan jenis fasilitas dari input pengguna secara otomatis. Studi oleh Badrudin & Sukenda (2025) mengimplementasikan NER berbasis BERT pada sistem chatbot *customer service* dan menunjukkan bahwa teknologi ini mampu memberikan respons yang cepat, akurat, dan situasional secara *real-time*, sekaligus mengurangi beban kerja staf dalam melayani pertanyaan pelanggan. Pendekatan serupa juga diterapkan oleh Putra et al. (2026) yang mengembangkan chatbot berbasis NLP menggunakan *framework* RASA dan IndoBERT pada layanan rekomendasi, dengan hasil F1-score mencapai 88,5% setelah optimasi dataset dan *hyperparameter*, membuktikan bahwa kombinasi model bahasa berbasis *transformer* dengan teknik klasifikasi *intent* mampu menghasilkan sistem rekomendasi yang interaktif dan efisien.

Letak geografis Pulau Jawa menjadi pertimbangan strategis tersendiri dalam pengembangan sistem rekomendasi akomodasi. Sebagai pulau dengan kepadatan penduduk tertinggi di Indonesia dengan jumlah penduduk mencapai 141 juta jiwa, sekaligus pusat aktivitas ekonomi dan pariwisata nasional, Pulau Jawa menawarkan ragam destinasi wisata yang tersebar di berbagai kota besar maupun daerah tujuan wisata alternatif. Tingginya volume pergerakan wisatawan domestik maupun mancanegara di wilayah ini menghasilkan permintaan informasi akomodasi yang sangat dinamis dan beragam. Kondisi tersebut mendorong kebutuhan akan sistem yang mampu menyajikan informasi secara cepat dan mampu menyesuaikan rekomendasi berdasarkan variasi preferensi pengguna, mulai dari kisaran harga yang terjangkau hingga alternatif opsi fasilitas di lokasi yang spesifik.

Penelitian sebelumnya sebagian besar masih terbatas pada lingkup data tertentu atau belum secara spesifik mengintegrasikan kebutuhan pengguna berdasarkan berbagai lokasi, bujet, dan fasilitas di Pulau Jawa. Oleh karena itu, masih terdapat kesenjangan antara kondisi ideal, yaitu sistem rekomendasi yang personal, cepat, dan interaktif, dengan kondisi aktual di mana pengguna masih melakukan pencarian manual melalui berbagai *platform*.

Penelitian ini mengusulkan pengembangan chatbot rekomendasi hotel di Pulau Jawa berbasis NLP yang mampu memahami kebutuhan pengguna melalui percakapan bahasa alami. Kebaruan penelitian ini terletak pada integrasi teknik NER dan *Cosine Similarity* untuk menghasilkan rekomendasi akomodasi yang signifikan berdasarkan prioritas fasilitas dan lokasi yang diinginkan pengguna secara *real-time*.

KAJIAN TEORI

AKOMODASI HOTEL DALAM INDUSTRI PARIWISATA

Akomodasi hotel dan motel merupakan salah satu komponen penting dalam sektor industri pariwisata yang berfungsi sebagai tempat tinggal sementara bagi wisatawan selama melakukan perjalanan. Akomodasi ini menyediakan berbagai pilihan layanan, fasilitas, serta rentang harga yang beragam sehingga dapat menyesuaikan dengan kebutuhan pengguna. Seiring dengan meningkatnya aktivitas wisata, jumlah akomodasi yang tersedia juga mengalami peningkatan. Hal ini menyebabkan wisatawan dihadapkan pada banyak pilihan yang harus dibandingkan secara manual (Gultom et al., 2025).

Kajian pada bidang pariwisata menunjukkan bahwa meningkatnya jumlah destinasi dan layanan wisata membuat kebutuhan terhadap sistem informasi yang mampu memberikan rekomendasi secara cepat dan tepat menjadi semakin penting (Kusnanda et al., 2022). Informasi akomodasi yang tersebar di berbagai platform digital seringkali menyulitkan pengguna dalam mengambil keputusan karena harus melakukan pencarian dan perbandingan secara mandiri (Wulandari et al., 2025). Oleh karena itu, pemanfaatan teknologi digital sangat diperlukan untuk membantu pengguna

memperoleh rekomendasi akomodasi secara lebih efisien.

NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk memahami, mengolah, dan merespons bahasa manusia dalam bentuk teks maupun suara (Da Costa et al., 2025). NLP banyak digunakan dalam sistem yang membutuhkan interaksi langsung dengan pengguna, seperti chatbot dan sistem rekomendasi.

Penelitian oleh Simorangkir et al. (2024) menunjukkan bahwa chatbot berbasis NLP mampu memahami pertanyaan pengguna dan memberikan tanggapan secara otomatis. Selain itu, NLP juga dapat dimanfaatkan dalam sistem rekomendasi berbasis teks dengan cara menganalisis input berupa ulasan pengguna untuk menghasilkan rekomendasi yang lebih substansial.

Penelitian yang dilakukan oleh Erlangga & Sari (2024) menunjukkan bahwa penerapan NLP dalam sistem rekomendasi dapat membantu memahami kebutuhan pengguna berdasarkan data teks yang diberikan. Hasil penelitian tersebut menunjukkan bahwa sistem berbasis NLP mampu memberikan pertimbangan dengan tingkat akurasi dan signifikansi yang tinggi, sehingga dapat membantu pengguna dalam mengambil keputusan secara lebih tepat.

CHATBOT DALAM SISTEM INFORMASI PARIWISATA

Chatbot merupakan program komputer yang dirancang untuk berkomunikasi dengan *user* melalui percakapan otomatis menggunakan teknologi NLP. Chatbot mampu meniru interaksi manusia sehingga pengguna dapat memperoleh informasi dengan cara yang lebih mudah dan cepat.

Penelitian oleh Kusumaningtyas et al. (2023) menunjukkan bahwa chatbot dapat meningkatkan efektivitas layanan data serta memberikan respons yang lebih cepat dan akurat dibandingkan sistem konvensional. Dalam sektor pariwisata, chatbot juga mampu meningkatkan pengalaman *user* melalui interaksi yang lebih dipersonalisasi dan dua arah.

Selain itu, penelitian oleh Young & Tukino (2025) menunjukkan bahwa chatbot berbasis NLP dapat membantu pengguna dalam memilih produk atau layanan dengan lebih efisien. Chatbot mampu

memahami kebutuhan pengguna berdasarkan input teks, kemudian memberikan *advice* sesuai dengan personalisasi seperti harga, kebutuhan, dan kategori tertentu. Hal ini membuat proses penggalan informasi menjadi lebih cepat dan tidak perlu dilakukan secara manual.

Penelitian lain oleh Nurhaliza et al. (2025) juga menunjukkan bahwa chatbot berbasis AI dengan pendekatan NLP mampu memahami maksud pengguna (*intent*) dan memberikan rekomendasi yang sesuai dengan kebutuhan. Chatbot ini terbukti dapat meningkatkan ketepatan dalam pencarian informasi serta membantu pengguna dalam mengambil keputusan yang cepat dan tepat.

NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition (NER) merupakan metode pada NLP yang digunakan untuk mengidentifikasi serta mengekstraksi informasi penting dari *unstructured text*. Entitas yang dikenali dapat berupa nama lokasi, harga, fasilitas, maupun kategori tertentu yang berhubungan dengan kebutuhan *user*.

Dalam penelitian terkait chatbot berbasis NLP yang dilakukan oleh Kusumaningtyas et al. (2023), metode NER digunakan untuk mengenali informasi utama dari pertanyaan pengguna sehingga sistem dapat memahami permintaan secara terkhusus. Penerapan NER memungkinkan proses ekstraksi preferensi pengguna seperti lokasi tujuan, anggaran, dan fasilitas akomodasi dilakukan secara otomatis sehingga meningkatkan akurasi sistem rekomendasi.

COSINE SIMILARITY

Cosine Similarity merupakan metode pengukuran kemiripan teks yang menghitung kesamaan antara dua representasi data berbasis vektor. Metode ini banyak digunakan dalam sistem pencarian informasi dan rekomendasi karena mampu menentukan tingkat kesesuaian antara kebutuhan pengguna dan data yang tersedia. Penelitian oleh Falah & Suryawan (2022) menunjukkan bahwa metode *Cosine Similarity* dapat digunakan untuk menghitung tingkat kemiripan antara data masukan pengguna dengan data dalam sistem rekomendasi sehingga menghasilkan rekomendasi yang sesuai dengan kebutuhan pengguna.

Untuk menghitung tingkat kemiripan antara dua vektor, digunakan rumus *Cosine Similarity* sebagai berikut:

$$\text{Cosine Similarity}(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

Nilai *Cosine Similarity* berada pada rentang 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan tingkat kemiripan yang tinggi antara dua vektor.

METODE

Penelitian ini menggunakan pendekatan eksperimen dalam pengembangan sistem berbasis kecerdasan buatan, yaitu chatbot rekomendasi akomodasi hotel berbasis NLP. Sistem dirancang untuk memahami kebutuhan pengguna melalui bahasa alami dan menghasilkan rekomendasi hotel yang sesuai dengan preferensi pengguna.

RANCANGAN PENELITIAN

Rancangan penelitian dilakukan melalui beberapa tahapan, yaitu pengumpulan data, pra-pemrosesan data, implementasi metode NLP, ekstraksi informasi menggunakan NER, perhitungan tingkat kemiripan menggunakan *Cosine Similarity*, integrasi sistem chatbot, serta evaluasi kinerja sistem. Tahapan ini disusun secara tertata untuk menghasilkan sistem rekomendasi yang interaktif dan cermat.

POPULASI DAN SAMPEL (SASARAN PENELITIAN)

Populasi dalam penelitian ini merupakan seluruh data akomodasi hotel yang tersedia di Indonesia. Sampel yang digunakan berupa dataset akomodasi hotel di wilayah Pulau Jawa yang diperoleh dari platform Kaggle. Dataset tersebut berisi informasi penting seperti nama hotel, lokasi, harga, fasilitas, dan deskripsi akomodasi yang digunakan sebagai dasar dalam proses rekomendasi.

TEKNIK PENGUMPULAN DATA

Teknik pengumpulan data dalam penelitian ini menggunakan data sekunder yang diperoleh dari dataset publik Kaggle. Data kemudian diseleksi dan disesuaikan dengan kebutuhan penelitian, yaitu hanya mengambil data akomodasi hotel pada Pulau Jawa. Selain itu, data masukan pengguna diperoleh dalam bentuk teks percakapan (input chatbot) yang

digunakan untuk menguji performa sistem dalam memahami kebutuhan pengguna.

Untuk memberikan gambaran mengenai data yang digunakan, contoh dataset akomodasi hotel ditampilkan pada Gambar 1.

id	hotel_name	property_type	city	address	facilities	rating	min_price	max_price	latitude	longitude	test
0	Maanvillan at Dharmaswathi Sundaras	Hotel	Gubung, Sundaras	Jl. Dharmaswathi No. 195, Map Gubung, Kota SR	[WiFi, Parking, Restaurant, Roomservice, 24]	4.2	1175000	356500	-7.267435	112.777124	Maanvillan at Dharmaswathi Sundaras Hotel [
1	O Canal Hotel Sundaras	Hotel	Subuhan, Sundaras	Jl. Gandra No. 8, Gandra Kcc Subuhan, Kota S	[NoOneside CARP, WiFi, Parking, Restaurant]	4.2	398075	197000	-7.247228	112.739480	O Canal Hotel Sundaras Hotel [NoOneside CARP]
2	Chakra City Hotel	Hotel	Subuhan, Sundaras	Subuhan, Sundaras City, East Java 60174, Indon	[NoOneside CARP, WiFi, CLEAN, WiFi, Pnc]	4.4	375000	270000	-7.255880	112.738886	Chakra City Hotel Hotel [NoOneside CARP, Y
3	G Sulas Hotel By AMITHA	Hotel	Gubung, Sundaras	Jl. Raya Gubung No.43, Gubung, Kcc, Gubung, Ki	[NoOneside CARP, WiFi, CLEAN, WiFi, Pnc]	3.9	188000	287000	-7.279623	112.747495	G Sulas Hotel By AMITHA hotel [NoOneside CA
4	Royal Regatta Candara Banyuw Royal Singos	Hotel	Togalan, Sundaras	Jl. Kumbia Pda Man, Dapad No.1, Togalan, K	[CLEAN, WiFi, Pnc]	4.2	1300000	380000	-7.286379	112.736621	Royal Regatta Candara Banyuw Royal Singos

Gambar 1. Cuplikan 5 data teratas akomodasi hotel dari kaggle

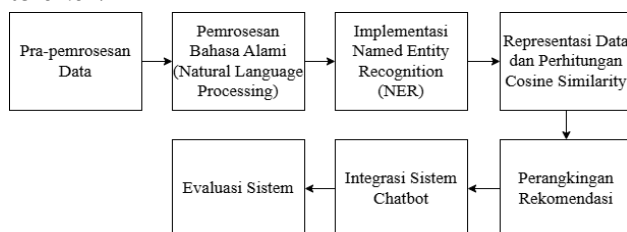
Berdasarkan Gambar 1, dataset yang digunakan terdiri dari beberapa atribut penting seperti nama hotel, lokasi, harga, dan fasilitas. Data tersebut menjadi dasar dalam proses pengolahan teks, mulai dari tahap *preprocessing* hingga pembobotan, yang selanjutnya digunakan dalam sistem rekomendasi chatbot.

PENGEMBANGAN INSTRUMEN

Instrumen yang digunakan dalam penelitian ini meliputi: (1) Dataset hotel berbasis *spreadsheet* (Microsoft Excel) sebagai basis data sistem; (2) Bahasa pemrograman Python untuk pengembangan sistem; (3) Library NLP seperti spaCy dan NLTK; (4) Library scikit-learn untuk perhitungan *Cosine Similarity*; (5) Sistem penyimpanan data berbasis *file* (Excel); (6) Antarmuka chatbot berbasis web.

TEKNIK ANALISIS DATA

Gambar 2 memperlihatkan teknik analisis data. Setiap tahap memiliki indikator capaian yang terukur.



Gambar 2. Teknik Analisis Data

Analisis data dilakukan melalui beberapa tahapan berikut:

1. Pra-pemrosesan Data (*Preprocessing*)

Dataset hotel yang diperoleh dari Kaggle diolah dalam format *spreadsheet* (Excel). Tahapan *preprocessing* meliputi *case folding*, tokenisasi, penghapusan *stopwords*, dan

normalisasi teks untuk menghasilkan data yang bersih dan siap diproses.

2. Pemrosesan Bahasa Alami (NLP)

Sistem chatbot memproses input pengguna dalam bentuk teks untuk memahami maksud dan *framing* pertanyaan menggunakan teknik NLP.

3. Named Entity Recognition (NER)

Metode NER digunakan untuk mengekstraksi informasi penting dari input pengguna, seperti lokasi, harga, dan fasilitas yang diinginkan. Proses ini membantu sistem dalam memahami keinginan pengguna secara spesifik.

4. Representasi Data dan Perhitungan Cosine Similarity

Data teks diubah menjadi representasi numerik menggunakan metode TF-IDF. Selanjutnya, dilakukan perhitungan *Cosine Similarity* untuk mengukur tingkat kesamaan antara preferensi pengguna dan data hotel. Nilai kemiripan yang lebih tinggi menunjukkan tingkat kesesuaian yang lebih baik.

5. Perangkingan Rekomendasi

Hasil perhitungan *Cosine Similarity* digunakan untuk mengurutkan data hotel sehingga diperoleh rekomendasi berdasarkan nilai kemiripan tertinggi.

6. Integrasi Sistem Chatbot

Seluruh proses diintegrasikan ke dalam sistem chatbot berbasis *web* sehingga pengguna dapat berinteraksi secara langsung untuk mendapatkan rekomendasi hotel.

7. Evaluasi Sistem

Evaluasi dilakukan untuk mengukur: (1) Kemampuan sistem dalam mengekstraksi entitas (akurasi NER); (2) Kesesuaian hasil rekomendasi; (3) Tanggapan sistem terhadap berbagai variasi input pengguna.

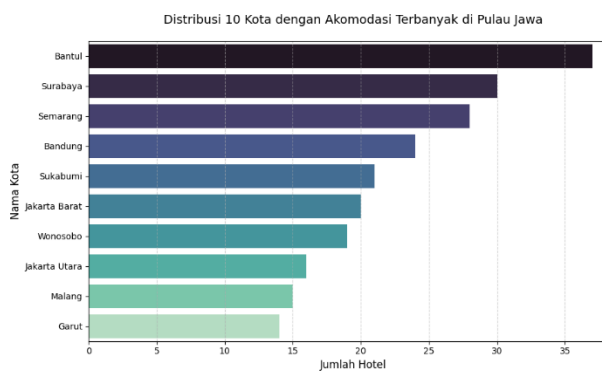
HASIL DAN PEMBAHASAN

SEBARAN HOTEL DI PULAU JAWA

Penelitian ini mengolah dataset hotel yang mencakup wilayah strategis di Pulau Jawa. Hasil pemetaan data menunjukkan variasi properti yang kaya, mulai dari hotel bintang lima hingga penginapan *budget*.

Tabel 1. Top 10 Kabupaten/Kota dengan Hotel Terbanyak di Pulau Jawa

Kabupaten/Kota	Frekuensi
Bantul	37
Surabaya	30
Semarang	28
Bandung	24
Sukabumi	21
Jakarta Barat	20
Wonosobo	19
Jakarta Utara	16
Malang	15
Garut	14



Gambar 3. 10 Kabupaten/Kota dengan Hotel Terbanyak di Pulau Jawa

Berdasarkan Tabel 1 dan Gambar 3, ditemukan fakta menarik bahwa Kabupaten Bantul memiliki frekuensi akomodasi tertinggi dalam dataset dengan jumlah 37 unit, diikuti oleh Kota Surabaya (30 unit), dan Kota Semarang (28 unit). Keberagaman persebaran memberikan justifikasi terhadap perlunya sistem rekomendasi berbasis chatbot.

Wilayah seperti Bantul, Sukabumi, dan Wonosobo yang masuk dalam daftar sepuluh besar menunjukkan bahwa pilihan akomodasi tidak hanya terkonsentrasi di kota metropolitan besar. Bagi wisatawan, banyaknya pilihan di wilayah wisata kabupaten sering kali menimbulkan kesulitan dalam menentukan akomodasi yang sesuai dengan fasilitas spesifik. Dengan demikian, kemampuan chatbot inapinep.id dalam memproses *query* deskriptif menjadi solusi mendasar untuk melakukan penyaringan data secara otomatis dan perseptif pada wilayah-wilayah dengan kepadatan penginapan yang tinggi tersebut.

EVALUASI TAHAPAN PREPROCESSING TEKS

Tabel 2. Transformasi Kata pada Tahapan Preprocessing

Tahap	Hasil Transformasi
Data Mentah	MaxoneHotels at Dharmahusada Surabaya, Hotel, "Gubeng, Surabaya", "[Wi-Fi, 'Parkir', 'Restoran', 'Resepsionis 24 Jam', 'Spa', 'Lift', 'AC', 'Fasilitas Rapat', ... 'Layanan Laundry/Dry Cleaning']"
Case Folding & Cleaning	maxonehotels dharmahusada surabaya hotel gubeng surabaya wifi parkir restoran resepsionis jam spa lift ac fasilitas rapat fasilitas ramah difabel ruang merokok lounge wifi publik ruang merokok khusus layanan laundry dry cleaning
Normalisasi Sinonim	maxonehotels dharmahusada surabaya hotel gubeng surabaya wifi parkir restoran resepsionis jam spa lift ac fasilitas rapat fasilitas ramah difabel ruang merokok lounge wifi publik ruang merokok khusus laundry
Stopword Removal	maxonehotels dharmahusada surabaya hotel gubeng surabaya wifi parkir restoran resepsionis jam spa lift ac fasilitas rapat fasilitas ramah difabel ruang merokok lounge wifi publik ruang merokok khusus laundry

Berdasarkan Tabel 2, terlihat bahwa proses *cleaning* berhasil mereduksi karakter seperti tanda petik, kurung siku, dan koma yang dapat mengganggu proses pembobotan kata. Normalisasi sinonim efektif untuk memetakan istilah-istilah yang memiliki makna serupa namun ditulis berbeda, contohnya mengubah "Layanan Laundry/Dry Cleaning" menjadi "laundry" seperti pada Tabel 3.

Tabel 3. Pemetaan Sinonim dan Pembersihan Frasa

Istilah Asli (Pattern)	Hasil Normalisasi
wi-fi, wi fi, wifi gratis	wifi
swimming pool, pool, kolam	kolam renang
laundry service, dry cleaning	laundry
restaurant, resto	restoran
television, tv	televisi

Pemetaan sinonim seperti pada Tabel 3 memungkinkan sistem untuk melakukan pencocokan semantik yang lebih akurat tanpa harus mengubah bentuk dasar kata melalui *stemming*. Keputusan untuk tidak menggunakan tahap *stemming* pada penelitian ini juga didasarkan pada karakteristik dataset hotel yang didominasi oleh kata benda (fasilitas dan lokasi) yang sudah merupakan kata dasar. Dengan menghindari *stemming*, sistem dapat mempertahankan integritas istilah teknis dan nama brand hotel secara utuh, sehingga meminimalisir risiko hilangnya makna kata (*over-stemming*) yang sering terjadi pada algoritma *stemming* Bahasa Indonesia.

HASIL EKSTRAKSI FITUR DAN REPRESENTASI VISUAL

Setelah melalui tahapan normalisasi dan penghapusan redundansi, dilakukan ekstraksi fitur untuk mengidentifikasi kata kunci yang paling representatif dalam dataset. Ekstraksi fitur ini bertujuan untuk memetakan karakteristik setiap akomodasi ke dalam ruang vektor yang nantinya akan diproses oleh algoritma *Cosine Similarity*. Hasil representasi fitur divisualisasikan melalui *WordCloud* pada Gambar 4.



Gambar 4. WordCloud

Berdasarkan Gambar 4, dapat diamati bahwa kata-kata seperti "wifi", "ac", "parkir", "resepsionis" dan "restoran" muncul dengan ukuran tipografi yang paling dominan. Hal ini mengonfirmasi bahwa fitur-fitur tersebut merupakan fasilitas standar dan paling konsisten yang ditawarkan oleh industri perhotelan di Pulau Jawa. Dominansi kata-kata ini sangat penting bagi mekanisme *chatbot*, karena asisten virtual harus mampu mengenali parameter dasar ini dengan akurasi tinggi sebelum memproses permintaan fasilitas yang lebih spesifik.

Hal yang paling signifikan dari hasil ekstraksi fitur ini adalah keberhasilan dalam deduplikasi kata tingkat dokumen. Pada observasi awal, ditemukan

bahwa frekuensi kata tertentu sangat bias karena pengulangan istilah yang sama dalam satu baris data hotel. Dengan diterapkannya pembersihan redundansi, Gambar 4 menampilkan sebaran fitur yang lebih seimbang, di mana kata kunci pendukung seperti "lift", "spa", "rapat", dan "laundry" mulai terlihat dengan porsi yang proporsional.

Selain itu, representasi lokasi seperti "Jakarta", "Surabaya", dan "Semarang" juga muncul sebagai fitur pembeda. Selain menangkap aspek fasilitas fisik, sistem juga menangkap aspek geografis yang menjadi salah satu pertimbangan utama pengguna dalam mencari akomodasi melalui chatbot inapinep.id. Kejelasan fitur yang terekstrak menjamin bahwa pembobotan pada tahap TF-IDF nantinya akan lebih akurat dan meminimalisir kesalahan rekomendasi akibat bias frekuensi kata.

ANALISIS PEMBOBOTAN FITUR DENGAN TF-IDF

Setelah melalui tahap ekstraksi fitur, setiap kata kunci diberikan bobot menggunakan metode TF-IDF. Tahap ini bertujuan untuk menentukan tingkat kepentingan suatu kata. Hasil pembobotan pada lima sampel hotel dengan fitur-fitur pilihan disajikan pada Tabel 4.

Tabel 4. Nilai TF-IDF pada Fitur Akomodasi

Nama Hotel	wifi	kola m renan g	spa	laundr y	rapat
maxonehot els	0,0618	0,0000	0,2526	0,3382	0,1460
d carol hotel	0,0813	0,0000	0,0000	0,0000	0,1922
choice city hotel	0,0557	0,0000	0,2278	0,0000	0,1317
g suites hotel by amithya	0,0417	0,1362	0,0000	0,2284	0,0986
royal regantris cendana	0,0516	0,0000	0,2110	0,0000	0,1220

Analisis terhadap Tabel 4 menunjukkan bahwa fitur "wifi" secara konsisten memiliki nilai bobot yang rendah di seluruh sampel hotel (berada di rentang antara 0,04 hingga 0,08). Hal ini dikarenakan fasilitas wifi merupakan fitur yang bersifat umum dan muncul di hampir semua dokumen, sehingga nilai IDF-nya menjadi rendah. Sebaliknya, fitur yang lebih spesifik seperti "spa" pada MaxoneHotels memiliki

bobot yang signifikan sebesar 0,2526, dan fitur "laundry" mencapai 0,3382.

Perbedaan nilai bobot ini membuktikan bahwa algoritma TF-IDF berhasil melakukan diskriminasi fitur dengan baik. Fitur yang unik dan jarang muncul di hotel lain diberikan bobot yang lebih tinggi, sehingga menjadi identitas utama hotel tersebut dalam ruang vektor. Prosedur ini memungkinkan sistem untuk memprioritaskan hotel yang memiliki kesesuaian fitur paling spesifik sesuai dengan permintaan pengguna.

UJI SIGNIFIKANSI FITUR DENGAN CHI-SQUARE

Penguatan dasar pemilihan fitur dalam sistem rekomendasi hotel ini dilakukan melalui uji statistik *Chi-Square*. Tujuan pengujian ini adalah untuk mengidentifikasi keterkaitan antara fitur tekstual dengan label kepuasan pengguna (rating). Hasil sepuluh fitur dengan skor tertinggi disajikan pada Tabel 5.

Tabel 5. Fitur Paling Signifikan Berdasarkan Skor *Chi-Square*

Fitur	Chi-Square	p-value
lift	7,2461	0,0071
sari	3,3613	0,0667
semarang	1,9028	0,1677
bandara	1,8233	0,1769
jemput	1,6659	0,1967
teras	1,4846	0,2230
taman	1,4632	0,2264
surabaya	1,3357	0,2477
rapat	1,3304	0,2487
penyimpanan_barang	1,2661	0,2604

Berdasarkan hasil uji *Chi-Square* pada Tabel 5, fitur "lift" muncul sebagai fitur yang paling signifikan dengan skor 7,2461 dan $p - value$ sebesar 0,0071. Mengingat nilai $p < 0,05$, maka dapat disimpulkan bahwa keberadaan fasilitas lift memiliki hubungan yang sangat signifikan terhadap klasifikasi rating hotel. Hal ini menunjukkan bahwa bagi wisatawan di Pulau Jawa, ketersediaan lift merupakan infrastruktur wajib dan genting yang secara langsung memengaruhi persepsi kualitas dan kenyamanan akomodasi, terutama pada bangunan hotel bertingkat.

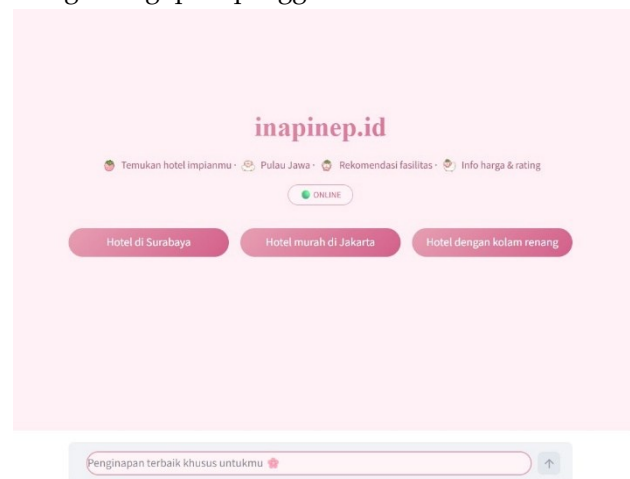
Fitur lain seperti "sari" (yang merujuk pada bagian dari nama daerah atau nama hotel spesifik) memiliki

$p - value$ sebesar 0,0667. Walaupun angka ini berada sedikit di atas ambang batas signifikansi 0,05, fitur ini tetap menunjukkan kecenderungan pengaruh yang lebih kuat dibandingkan fitur lokasi lainnya seperti "semarang" ($p = 0,1677$) atau "surabaya" ($p = 0,2477$).

Munculnya fitur layanan transportasi seperti "bandara" dan "jemput" dalam daftar sepuluh besar juga memberikan *insight* tambahan. Meskipun secara statistik tidak signifikan pada taraf 5%, kehadiran fitur ini menunjukkan bahwa aksesibilitas dan layanan penjemputan tetap menjadi variabel yang diperhatikan oleh pengguna saat memberikan penilaian. Temuan ini memberikan landasan bagi sistem chatbot inapinep.id untuk memberikan bobot perhatian lebih pada fitur infrastruktur utama seperti lift saat melakukan proses filterisasi, karena fitur tersebut terbukti menjadi pembeda antara hotel dengan rating tinggi dan rendah dalam dataset ini.

IMPLEMENTASI ANTARMUKA CHATBOT INAPINEP.ID

Chatbot sistem rekomendasi hotel yang dikembangkan diintegrasikan ke dalam antarmuka berbasis web menggunakan *framework* Streamlit. Pemilihan Streamlit didasarkan pada efisiensinya dalam menangani *pipeline* data sains secara *real-time*. Antarmuka inapinep.id dirancang dengan konsep *clean* dan minimalis guna memastikan kemudahan navigasi bagi para pengguna.



Gambar 5. Tampilan Antarmuka Awal Chatbot inapinep.id

Gambar 5 menunjukkan halaman beranda sistem chatbot di mana pengguna disambut oleh saran kueri interaksi. Setiap pesan yang dikirimkan pengguna akan memicu fungsi pemrosesan di *back-end* yang melakukan transformasi teks secara instan sesuai

dengan model TF-IDF yang telah dilatih. Penggunaan ngrok sebagai *tunneling* memungkinkan model yang berada di lingkungan *cloud computing* dapat diakses secara publik, sehingga memberikan fleksibilitas dalam pengujian aksesibilitas sistem.

SIMULASI RELEVANSI HASIL REKOMENDASI

Evaluasi kinerja model dilakukan melalui simulasi *query* deskriptif untuk menguji ketajaman algoritma *Cosine Similarity*. Sistem memetakan *query* pengguna ke dalam ruang vektor dan mencari dokumen hotel dengan sudut kosinus terkecil.

Pada percobaan simulasi ini, peneliti menguji tiga skenario kueri deskriptif yang berbeda untuk melihat tingkat sensitivitas algoritma *Cosine Similarity* dalam memproses kueri kombinasi antara fitur lokasi dan fasilitas. Skenario pertama menggunakan kueri "*hotel murah di Surabaya ada fasilitas kolam renang*". Hasil lima rekomendasi hotel teratas disajikan pada Tabel 6.

Tabel 6. Hasil Rekomendasi Skenario 1

Nama Hotel	Kab/Kota	Skor Similaritas
Bisanta Bidakara Surabaya	Surabaya	0,5894
Life Taman Melati Surabaya	Surabaya	0,5163
Prime Royal Hotel Surabaya	Surabaya	0,5160
Hotel Great Diponegoro	Surabaya	0,4740
Halogen Hotel Airport	Surabaya	0,4740

Berdasarkan Tabel 6, sistem menunjukkan akurasi geografis yang sempurna pada wilayah Surabaya. Hotel Bisanta Bidakara Surabaya menempati urutan pertama karena memiliki kecocokan vektor yang kuat pada fitur lokasi, fasilitas kolam renang, dan label harga yang kompetitif pada metadata. Rentang skor 0,47 – 0,58 menunjukkan bahwa model TF-IDF berhasil melakukan diskriminasi fitur dengan baik pada wilayah dengan densitas hotel tinggi.

Skenario kedua menggunakan kueri: "*rekomendasi hotel di Jakarta yang ada fasilitas ruang rapat*". Hasil lima rekomendasi teratas ditampilkan pada Tabel 7.

Tabel 7. Hasil Rekomendasi Skenario 2

Nama Hotel	Kab/Kota	Skor Similaritas
SPOT ON 2754 Miyagi	Sleman	0,6687

RedDoorz Resort Darajat	Garut	0,6407
Penginapan Putra Darma	Gunung Kidul	0,6215
OYO Life 3003 Wisma H.	Gunung Kidul	0,6157
Alzara Hotel Syariah	Gunung Kidul	0,6157

Pada Tabel 7, ditemukan fenomena menarik di mana sistem memberikan rekomendasi hotel yang berada di luar wilayah Jakarta. Hal ini disebabkan oleh fitur "ruang rapat" pada hotel-hotel tersebut memiliki bobot nilai TF-IDF yang sangat tinggi dalam korpus data, mengalahkan bobot fitur lokasi "Jakarta". Hal ini membuktikan bahwa algoritma sangat sensitif terhadap fasilitas spesifik yang jarang muncul (langka). Meskipun terjadi disparitas lokasi, secara fungsional sistem tetap memberikan opsi hotel yang memiliki fasilitas sesuai dengan permintaan pengguna dengan tingkat kemiripan vektor yang tinggi ($> 0,60$).

Skenario ketiga menggunakan kueri "*hotel di Semarang yang dekat dengan restoran*". Hasil kelima rekomendasi teratas ditunjukkan pada Tabel 8.

Tabel 8. Hasil Rekomendasi Skenario 3

Nama Hotel	Kab/Kota	Skor Similaritas
Hotel Catur	Magelang	0,5911
Pakde air cafe and homestay	Magelang	0,5230
RedDoorz Patung Kuda Undip	Semarang	0,5095
Smart Budget Hotel	Semarang	0,4986
OYO 2393 Omahe Nawung	Magelang	0,4979

Tabel 8 menunjukkan hasil yang cukup kompetitif antara wilayah Magelang dan Semarang. Munculnya RedDoorz Patung Kuda Undip di peringkat ketiga memvalidasi bahwasanya sistem masih mampu menangkap fitur lokasi Semarang. Namun, adanya dominasi hotel di Magelang pada peringkat satu dan dua disebabkan oleh kata kunci "restoran" yang muncul secara repetitif dalam deskripsi hotel tersebut. Hasil perolehan ini memberikan wawasan penting bagi pengembangan sistem selanjutnya untuk melakukan *fine-tuning* pada bobot fitur lokasi agar memiliki nilai koefisien yang lebih besar dibandingkan fitur fasilitas umum.

PENANGANAN KUERI DI LUAR TOPIK

Chatbot inapinep.id merupakan asisten virtual pintar yang dilengkapi dengan logika validasi input. Pengujian dilakukan dengan memasukkan pertanyaan yang tidak relevan dengan topik (*out of topic*) rekomendasi hotel di Pulau Jawa.



Gambar 6. Respon Chatbot inapinep.id Terhadap Input yang Tidak Relevan

Sebagaimana terlihat pada Gambar 6, sistem secara cerdas mendeteksi bahwa kueri "*dimana pantai selatan itu*" tidak memiliki irisan fitur dengan korpus hotel. Alih-alih memberikan rekomendasi acak, sistem chatbot inapinep.id memberikan respon penolakan yang kemudian mengarahkan pengguna untuk kembali ke topik perhotelan. Mekanisme *guardrails* ini sangat penting untuk menjaga integritas sistem dan tingkat kepercayaan pengguna.

PENUTUP

SIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa pengembangan chatbot rekomendasi hotel berbasis *Natural Language Processing* (NLP) dengan integrasi *Named Entity Recognition* (NER) dan *Cosine Similarity* mampu memahami kebutuhan pengguna serta menghasilkan rekomendasi akomodasi yang relevan. Proses *preprocessing* dan pembobotan TF-IDF terbukti efektif membentuk representasi fitur dalam membedakan karakteristik tiap hotel secara akurat.

Hasil pengujian menunjukkan bahwa metode *Cosine Similarity* mampu mengukur tingkat kemiripan antara input pengguna dan data hotel dengan baik, dengan nilai similaritas yang berada pada rentang 0,47 – 0,67 pada berbagai skenario pengujian. Nilai tertinggi mencapai 0,6687, yang menunjukkan tingkat relevansi yang cukup tinggi antara *query* pengguna dan rekomendasi yang dihasilkan oleh sistem.

Pembobotan fitur dengan TF-IDF menunjukkan bahwa fitur umum seperti wifi memiliki bobot

rendah pada rentang 0,04 – 0,08, sedangkan fitur yang lebih spesifik seperti spa dan laundry memiliki bobot yang lebih tinggi, masing-masing mencapai 0,2526 dan 0,3382. Hal ini membuktikan bahwa metode TF-IDF mampu melakukan diskriminasi fitur dengan baik berdasarkan tingkat kepentingannya dalam dataset.

Berdasarkan uji *Chi-Square*, fitur lift memiliki nilai signifikansi tertinggi dengan skor 7,2461 dan *p – value* sebesar 0,0071 ($p < 0,05$), yang menunjukkan bahwa fitur tersebut memiliki pengaruh yang signifikan terhadap klasifikasi kualitas hotel. Temuan ini menjelaskan bahwa fasilitas tertentu memiliki kontribusi yang lebih besar dalam membedakan karakteristik hotel dalam sistem rekomendasi.

Namun, ditemukan bahwa sistem memiliki tendensi untuk lebih memprioritaskan fitur fasilitas dibandingkan fitur lokasi, sehingga pada suatu kondisi rekomendasi yang dihasilkan belum sepenuhnya sesuai dengan preferensi wilayah pengguna. Hal ini terlihat pada skenario pengujian kedua Tabel 7, di mana hotel dengan fasilitas yang lebih spesifik dapat mengungguli kecocokan lokasi. Hal ini menunjukkan bahwa keseimbangan pembobotan fitur menjadi aspek penting dalam meningkatkan kualitas sistem rekomendasi.

Dengan demikian, penelitian ini menegaskan bahwa kombinasi NLP, NER, TF-IDF, dan *Cosine Similarity* merupakan pendekatan yang efektif dalam pengembangan sistem rekomendasi hotel di Pulau Jawa berbasis chatbot. Hasil penelitian ini membuka peluang pengembangan lebih lanjut, khususnya pada optimalisasi pembobotan fitur lokasi dan fasilitas agar sistem menjadi lebih adaptif dan mampu memberikan rekomendasi hotel yang lebih presisi sesuai dengan kebutuhan pengguna.

SARAN

Penelitian selanjutnya dapat memperluas ruang lingkup penelitian. Penelitian ini berfokus pada pengembangan chatbot rekomendasi hotel di Pulau Jawa, sehingga penelitian selanjutnya dapat mempertimbangkan cakupan wilayah yang lebih luas agar sistem mampu memberikan rekomendasi yang lebih komprehensif.

Penelitian selanjutnya dapat mengembangkan metode yang digunakan, khususnya pada peningkatan pembobotan fitur lokasi dan fasilitas yang lebih seimbang. Pengembangan pada aspek

data dan fitur personalisasi pengguna juga dapat dipertimbangkan untuk meningkatkan akurasi dan relevansi hasil rekomendasi.

DAFTAR PUSTAKA

- Badrudin, D. R., & Sukenda. (2025). Implementasi Named Entity Recognition untuk Botchat Customer Service PT. Afbe Cahaya Kreatif. *Jurnal Algoritma*, 22(2), 103-114. <https://doi.org/10.33364/algoritma/v.22-2.2394>
- Chandra, A. A., Nathaniel, V., Satura, F. R., & Adhinata, F. D. (2022). Pengembangan Chatbot Informasi Mahasiswa Berbasis Telegram dengan Metode Natural Language Processing. *Jurnal ICTEE*, 3(1), 20-27.
- Da Costa, A., Maulindar, J. and Hartanti, D., 2025. Sistem Rekomendasi Pemilihan Hotel Di Area Surakarta Menggunakan Metode Knowledge Based Dengan Pendekatan Constrain Based. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, 7(1), pp.230-235.
- Erlangga, F., & Sari, I. P. (2024). Perancangan sistem untuk merekomendasikan produk skincare menggunakan metode Natural Language Processing (NLP). *PRINSIP: Portal Riset dan Inovasi Sistem Perangkat Lunak*, 2(4). <https://doi.org/10.59696/prinsip.v2i4.49>
- Falah, Z.F. and Suryawan, F., 2022. Recommendation system to propose final project supervisors using cosine similarity matrix. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, 8(2), pp.156-162.
- Firanti, M., 2026. Strategi Meningkatkan Pendapatan Nasional Melalui Pengembangan Sektor Pariwisata Di Indonesia. *Journal of Economics, Commerce, and Organization*, 1(1), pp.23-33.
- Gultom, H.T., Darmagama, D.A.F. and Ardinnansyah, A., 2025. DAMPAK IMPLEMENTASI CHATBOT TERHADAP EFISIENSI LAYANAN PELANGGAN DAN KEPUASAN KONSUMEN DALAM KONTEKS BISNIS DIGITAL (2019-2025). *Journal Education, Sociology and Law*, 1(4), pp.1639-1651.
- Hananto, K., 2023. Persepsi Tamu Terhadap Penggunaan Teknologi di Hotel-Hotel Yogyakarta Menuju Pariwisata Cerdas. *Metta: Jurnal Ilmu Multidisiplin*, 3(3), pp.328-339.
- Kusnanda, I.G.R., Sukarsa, I.M. and Susila, A.A.N.H., 2022. Perancangan Chatbot Hotel dengan Model Natural Language Processing Chatbot dan Button Based Chatbot. *Jurnal Ilmiah Teknologi dan Komputer*, 3(1), pp.711-720.
- Kusumaningtyas, E.M., Laurentino, E.R. and Barakbah, A.R., 2022. Responsive Chatbot Using Named Entity Recognition and Cosine Similarity. In *Proceedings of the 5th International Conference on Applied Science and Technology on Engineering Science (iCAST-ES)*.
- Nurhaliza, S., Bustami, & Aidilof, H. A. K. (2025). Implementasi chatbot AI untuk rekomendasi produk skincare menggunakan Natural Language Processing. *Jurnal Ilmiah Global Education*, 6(2), 294-303. <https://doi.org/10.55681/jige.v6i2.3802>
- Putra, S. M. W., Putra, M. A. P., & Purnama, I. N. (2026). Penerapan Chatbot Berbasis Natural Language Processing pada Website Tarumenyan sebagai Rekomendasi Konsep Prewedding Bali. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 10(2), 2803-2808.
- Simorangkir, A., Sihite, P.I., Kiareni, C.L., Priskila, R. and Pranatawijaya, V.H., 2024. Pemodelan Chatbot Rekomendasi Hotel Dengan Menggunakan Natural Language Processing. *Simtek: jurnal sistem informasi dan teknik komputer*, 9(1), pp.46-50.
- Wulandari, N., Wijaya, D.A., Wijayanto, N., Marimin, M., Rifani, B.Y. and Taufiq, M.L., 2025. Analisis Pengembangan Wisata Pantai Watu Karung Kabupaten Pacitan melalui Pendekatan 4A. *Gemawisata: Jurnal Ilmiah Pariwisata*, 21(2), pp.235-248.
- Young, F., & Tukino. (2025). Perancangan dan implementasi chatbot berbasis Natural Language Processing untuk rekomendasi laptop. *Jurnal Comasie*, 12(1). <http://ejournal.upbatam.ac.id/index.php/comasiejurnal>.