An Intelligent Multi-Source Information Retrieval System for Hospital Chain Comparison using LLM-based Analysis

Apaar Raina, Aadish Jain, Krishna Agarwal

Abstract

This paper presents an intelligent information retrieval system that aggregates multi-source data from online reviews and news articles to generate comparative evaluations of hospital chains. Our approach employs dynamic web scraping using Selenium to collect reviews from platforms including JustDial and MouthShut, along with news headlines from Google News. The collected data is processed through three Large Language Models (Llama, Claude, and Ollama) to perform sentiment analysis and generate comprehensive comparative reports across multiple dimensions including patient care quality, infrastructure, staff behavior, and reputation. Manual evaluation demonstrates coherent and insightful comparisons. The modular architecture enables easy adaptation to other domains.

Keywords: Information Retrieval, Web Scraping, Large Language Models, Sentiment Analysis, Healthcare Analytics

1 Introduction

1.1 Background and Motivation

Healthcare decision-making significantly impacts patient outcomes, yet information about healthcare providers is fragmented across multiple online platforms. Patients face challenges in comprehensively evaluating hospital chains due to scattered reviews, varying rating systems, and lack of integrated analysis combining patient experiences with media coverage [1]. Traditional hospital evaluation methods focus on clinical outcomes while overlooking crucial patient experience dimensions.

Current approaches suffer from: (1) fragmented information across platforms, (2) time-intensive manual research, (3) subjective individual biases, (4) lack of temporal context, and (5) limited comparative analysis tools. The emergence of Large Language Models (LLMs) offers new opportunities for automated text analysis and synthesis [5].

1.2 Contributions

Our main contributions include: (1) a multi-source data aggregation framework using Selenium for dynamic scraping, (2) LLM-based comparative analysis using three models (Llama, Claude, Ollama), (3) multi-dimensional evaluation across patient care, infrastructure, staff behavior, and reputation, (4) domain-agnostic modular architecture, and (5) qualitative validation methodology for LLM-generated reports.

2 Literature Review

2.1 Web Scraping and Data Collection

Mitchell introduced comprehensive web scraping approaches, emphasizing dynamic content extraction challenges [2]. Selenium WebDriver has emerged as powerful for JavaScript-rendered content. Glez-Peña et al. demonstrated that dynamic scraping is essential for modern web applications with client-side rendering [3].

2.2 Healthcare Sentiment Analysis

Greaves et al. analyzed patient feedback from online sources, demonstrating strong correlations between online ratings and official quality metrics [1]. Hao and Zhang proposed methods for mining healthcare reviews to extract patient experiences, emphasizing aspect-based sentiment analysis [4].

2.3 Large Language Models

Brown et al. introduced GPT-3, demonstrating few-shot learning capabilities for complex text analysis [5]. Touvron et al. presented Llama, an open-source LLM achieving competitive performance while being accessible to researchers [6]. Open-source models have democratized access to advanced NLP capabilities.

2.4 Multi-Source Aggregation

Zhang and Liu explored techniques for aggregating information from multiple sources, highlighting importance of source credibility and conflict resolution [7]. Dong et al. proposed knowledge fusion methods for integrating facts from sources with varying reliability [8].

2.5 Gap Analysis and Novelty

Existing research addresses individual components but gaps remain: (1) limited integration of reviews and news for entity comparison, (2) lack of LLM-based healthcare provider comparison systems, (3) insufficient multi-model LLM approaches, and (4) limited domain-agnostic frameworks.

Our project's novelty lies in combining real-time review scraping from multiple platforms with live news aggregation, and analysing the collected data using mul-

tiple LLMs working together for more reliable results.It can be applied to any field by simply updating the input URLs.The system also provides a clear reasoninguseing LLMs and uses Selenium to handle and extract data from websites that load content dynamically.

3 Objective

Primary Objective: Develop an intelligent IR system that automatically collects, processes, and synthesizes multi-source data to generate comprehensive hospital chain comparisons.

Secondary Objectives: (1) Implement robust Selenium-based web scraping for JustDial, MouthShut, and Google News, (2) integrate and compare three LLM models for analysis, (3) evaluate hospitals across patient care, infrastructure, staff behavior, and reputation, (4) generate human-readable comparative reports, (5) qualitatively assess LLM output coherence and accuracy, and (6) validate system modularity for domain adaptation.

4 Proposed Model

4.1 System Architecture

The system comprises 2 pipeline components:

Web Scraping Module: Dynamic Selenium-based scraper with multi-platform support, configurable URLs, and error handling. This saves the reviews and the news into json files to be given as input to the LLM.

LLM Analysis Engine: Works with Llama, Claude, and Ollama, uses prompt design for comparing results, and includes a framework to evaluate the data from multiple angles.

4.2 Data Flow

The sequential flow: (1) Input hospital names and URLs, (2) Selenium extracts reviews and news and stores them as json files (3) LLMs process the files with analysis prompts (4) outputs are saved as txt files.

5 Methodology

5.1 Web Scraping Implementation

We employ Selenium WebDriver for JavaScript-rendered content handling. The scraper implements browser automation, waits for dynamic loading, scrolls for infinite-scroll content, and extracts fully-rendered data.

5.2 LLM Integration

The system integrates three language models—*Llama*, *Claude*, and *Ollama*—each contributing different strengths such as detailed reasoning, sentiment understanding, and lightweight local inference. A structured prompt is used to guide the models to compare multiple hospitals across key parameters, including patient satisfaction, staff quality, hygiene, infrastructure, emergency response, treatment specialties, affordability, wait times, and recent news.

The prompt instructs the model to generate: (1) overall rankings, (2) parameter-wise comparisons with evidence from reviews, (3) specialization-based insights, (4) strengths and weaknesses for each hospital, (5) patient-focused recommendations, (6) improvement suggestions, and (7) critical warnings. The analysis is required to be detailed, data-grounded, and supported by direct quotes from reviews, ensuring a thorough multi-dimensional evaluation.

5.3 Multi-Factor Analysis

Evaluation dimensions: *Patient Care* (treatment effectiveness, wait times, communication), *Infrastructure* (cleanliness, equipment, emergency services), *Staff Behavior* (professionalism, responsiveness, empathy), and *Reputation* (media sentiment, awards, compliance).

5.4 Evaluation Methodology

Manual evaluation assesses: *Coherence* (logical flow, consistency), *Relevance* (alignment with reviews, citation accuracy), *Comprehensiveness* (factor coverage, balanced perspectives), and *Cross-Model Comparison* (consensus findings, divergent interpretations).

6 Experimentation and Results

6.1 Dataset Description

We collected review and news data for seven major hospital chains in Delhi: BLK, Medanta, Fortis, Max, Apollo, Narayana, and AIIMS. The dataset spans January–November 2024 and combines multi-platform user reviews with real-time news articles for each hospital. Table 1 summarizes the final dataset statistics.

Table 1: Dataset Statistics (Reviews and News per Hospital)

Hospital	Total Reviews	Total News	Avg. Length
BLK	79	41	775.66
Medanta	102	27	1067.97
Fortis	114	27	251.69
Max	206	54	410.83
Apollo	102	37	435.60
Narayana	110	46	471.09
AIIMS	102	50	162.30

Across all hospitals, review sources included JustDial and MouthShut. Most hospitals contained a balanced mix, with MouthShut providing longer, more narrative reviews, while JustDial reviews were generally shorter and more compact. Source distributions were as follows: BLK (66 MouthShut, 13 JustDial), Medanta (100 MouthShut, 2 JustDial), Fortis (114 JustDial), Max (100 MouthShut, 106 JustDial), Apollo (100 MouthShut, 2 JustDial), Narayana (100 MouthShut, 10 JustDial), and AIIMS (92 MouthShut, 10 JustDial).

The news corpus included coverage of a variety of event types, such as medical achievements, administrative updates, patient incidents, recognitions, and regulatory announcements. This blend of long-form reviews and topical news articles provides a rich foundation for multi-dimensional LLM-based analysis and cross-hospital comparison.

6.2 LLM Model Comparison

A comparative evaluation of the three LLMs—Llama 3.1 (70B), Claude 3.5 Haiku, and Ollama—was carried out by manually analyzing their generated hospital assessment reports.

Llama: Produced highly detailed explanations and covered multiple aspects of patient experience, hospital quality, and medical outcomes. It tended to be more verbose but offered strong contextual reasoning. Quality scores: 8.4/10 (coherence), 8.1/10 (relevance), 8.7/10 (depth of analysis).

Claude: Provided balanced summaries, captured subtle sentiment patterns, and demonstrated strong contextual understanding across hospitals. Its analysis was more concise while still maintaining nuance. Quality scores: 8.8/10 (coherence), 8.5/10 (relevance), 8.4/10 (depth of analysis).

Ollama: Generated fast, locally executed summaries suitable for privacy-focused environments. Its interpretations were more surface-level and occasionally missed deeper sentiment cues. Quality scores: 7.6/10 (coherence), 7.4/10 (relevance), 7.1/10 (depth of analysis).

6.3 Comparative Analysis Results

The consolidated LLM evaluation—spanning reviews, complaints, and news sentiment—identified clear differences across the seven hospital chains (AIIMS, Max, BLK, Narayana, Apollo, Fortis, Medanta).

AIIMS: Strong medical outcomes, expert doctors, and high trust among patients. Despite high patient load, overall sentiment remained positive, yielding an estimated rating of 4.3/5. Overall: Excellent.

Max: Clean facilities and professional staff contributed to a solid reputation, though complaints about billing and occasional staff behaviour reduced consistency. Estimated rating: 3.9/5. Overall: Good.

BLK: Mixed experiences—patients praised certain doctors but frequently reported administrative delays and room-allocation issues. Estimated rating: 3.6/5. Overall: Fair.

Narayana: Recognized for paediatric and cardiac care, though reviews indicated long waiting times and inconsistent staff responsiveness. Estimated rating: 3.7/5. Overall: Good for specialized care.

Apollo: Strong infrastructure and reputed doctors but recurring complaints around billing transparency and communication. Estimated rating: 3.8/5. Overall: Good.

Fortis: Increasing number of negative reviews centered on staff behaviour and high costs. Estimated rating: 3.4/5. Overall: Below average.

Medanta: Recurring concerns about rude behaviour and negligence overshadowed good facilities. Estimated rating: 3.2/5. Overall: Weak.

6.4 Cross-Model Consensus

Across models, a clear pattern emerged:

High agreement:

- AIIMS consistently ranked highest due to strong treatment outcomes and trust.
- Max and Apollo formed the next tier, with good facilities but administrative drawbacks.
- Medanta and Fortis showed repeated negative sentiment across all LLMs.
- Narayana was consistently recognized for cardiac and neonatal care.

Model differences:

- Llama emphasized quantitative reasoning (e.g., inferred scores, ranking stability).
- Claude focused on temporal sentiment changes, highlighting specific 2024 incidents and service trends.
- Ollama produced shorter, high-level assessments, creating less granularity in differentiating close-ranked hospitals.

6.5 Discussion

Using a multi-LLM pipeline provided a robust cross-validation mechanism for hospital quality assessment. Reviews offered patient-level insights, while news articles added broader context such as awards, controversies, and administrative events. The combination significantly improved stability of sentiment interpretation.

Each LLM contributed distinct strengths:

- **Llama**: Best for comprehensive and research-oriented summaries.
- **Claude**: Most reliable for balanced, nuanced comparisons.
- **Ollama**: Ideal for fast, offline, privacy-preserving analysis.

7 Conclusions and Limitations

7.1 Conclusions

This work demonstrates that automated hospital-chain comparison is feasible, scalable, and practically valuable when combining multi-source data aggregation with multi-LLM reasoning. The key conclusions are:

(1) **Feasibility of Automated Comparative Analysis:** The system successfully generates coherent, multi-dimensional evaluations of hospitals by integrating reviews and news data with LLM-based reasoning. This

replaces several hours of manual research with fully automated reports.

(2) **Value of Multi-Source Aggregation:** Using both patient reviews and news articles enabled analyses that capture not only individual patient experiences but also organizational events, awards, controversies, and systemic issues. This holistic view proved crucial for fair hospital comparison.

(3) **Complementary LLM Strengths:** The three models exhibited distinct benefits: Llama provided detailed, research-style explanations; Claude produced nuanced, balanced interpretations; Ollama delivered fast, local, privacy-preserving insight. Cross-model agreement improved robustness and reduced the chance of single-model bias.

(4) **Generalizable Architecture:** The modular design allowed rapid adaptation to different domains (e.g., restaurant comparisons on Zomato/Swiggy). Only URL and parameter configurations required modification, confirming strong domain-transfer capability.

(5) **Practical Impact:** The generated outputs—rankings, strengths, weaknesses, risk flags, and recommendations—can support patients in informed decision-making, help hospital administrators evaluate service gaps, and assist researchers studying healthcare perception trends.

7.2 Limitations

Despite its effectiveness, the system has several limitations that highlight opportunities for future improvements:

(1) **Scraping Robustness:** Platforms such as JustDial and MouthShut frequently triggered anti-bot mechanisms. These required retry logic and occasional manual intervention, limiting full automation.

(2) **Lack of Objective Ground Truth:** Comparative hospital quality lacks standard benchmarks. Model outputs were validated manually, which introduces personal judgment and evaluator bias. Future work should incorporate quantitative medical outcomes or expert-annotated datasets.

(3) **Temporal Skew:** Recent reviews are more visible and may disproportionately affect sentiment trends. A temporal weighting or decay model would reduce recency bias in long-term analysis.

(4) **LLM Hallucination Risk:** Models sometimes infer details not explicitly present in the data. More strict citation enforcement or retrieval-augmented generation may further reduce hallucination.

(5) **Computational and Financial Cost:** High-quality LLMs such as Claude incur significant cost when generating detailed reports for many hospitals. While Ollama reduces cost and preserves privacy, its analytical depth is lower.

(6) **Non-Continuous Data Updates:** Data refresh is triggered manually. Automated, scheduled scraping and analysis would ensure up-to-date evaluations aligned with real-world hospital performance changes.

(7) **Review Authenticity:** The system treats all reviews as genuine. Detection of fake, incentivized, or bot-generated reviews—known issues on Indian platforms—was not implemented.

(8) **Limited Human Evaluation:** Validation was performed on a small sample of reviews. Large-scale human evaluation or comparison with healthcare experts would strengthen confidence in the system's assessments.

8 References

References

[1] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L., 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, 15(11), p.e239.

[2] Mitchell, R., 2018. *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, Inc.

[3] Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F., 2014. Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), pp.788-797.

[4] Hao, H. and Zhang, K., 2016. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *Journal of Medical Internet Research*, 18(5), p.e108.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., et al., 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, pp.1877-1901.

[6] Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al., 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[7] Zhang, Y. and Liu, B., 2014. Multi-source information fusion based on rough set theory: A review. *Information Fusion*, 17, pp.59-71.

[8] Dong, X.L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S. and Zhang, W., 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9), pp.938-949.