Scaling Multilingual E-Commerce with GenAI: Content Generation and Sentiment Analysis for Low-Resource Languages

Kathakali Mitra *1 **Kathakali Mitra** *1 **Harsh Khanna** *2

*Equal contribution [1]*Department of Computer Science and Information Systems, BITS Pilani, Hyderabad Campus, India* [2]*Dell Technologies*. Correspondence to: Kathakali Mitra p20220104@hyderabad.bits-pilani.ac.in.

1. Introduction

The rapid rise of e-commerce has expanded global retail, but engaging users in low-resource languages remains a major challenge. Multilingual Large Language Models (MLLMs)—such as ChatGPT, Llama, Mistral, Phi-3, Gemini, and Gemma—offer strong text generation and sentiment analysis abilities, yet their performance in low-resource languages is limited by scarce data. This creates an important research opportunity in multilingual content generation and evaluation. This paper explores the use of MLLMs for multilingual content creation and sentiment analysis in low-resource Indic languages, focusing on Hindi and Bengali. The key contributions are:

- **Multilingual Content Augmentation:** The study generates Indic product information (titles, descriptions, features) using Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) prompting with MLLMs. It promotes inclusive access to e-commerce content in regional languages. A comparative evaluation is performed using an LLM-as-a-judge framework, assessing Accuracy, Relevance, Consistency, Completeness, and Cohesiveness.
- **Multilingual Sentiment Analysis:** MLLMs are used to analyze complex product reviews expressed in low-resource Indic languages. These reviews may contain mixed, cross-domain, ambiguous, sarcastic, or incomplete emotions. The models also produce percentage-based emotion distributions for such mixed-sentiment cases.

2. Methodology

This section presents two complementary methodologies. The first focuses on multilingual product content generation in low-resource Indic languages using RAG and CoT prompting with MLLMs. Output quality is assessed using an LLM-as-a-Judge framework, along with a comparative evaluation of leading MLLMs—ChatGPT-4, Gemini-1.5 Flash-8B, Llama-3.1-8B Instruct, and Mistral-8×7B Instruct.

The second methodology addresses multilingual sentiment analysis of complex product reviews involving mixed, ambivalent, cross-domain, incomplete, and sarcastic sentiments. Both approaches employ prompt-based strategies and utilize synthetic data generation to address data scarcity. The overall architecture is shown in Fig. 1.

- **Synthetic Data Generation**

  We employed a MLLM to generate synthetic e-commerce data, including product titles, features, descriptions, and brand information. Carefully designed prompts were used to mimic real-world product data, ensuring relevance, accuracy, and linguistic diversity.

  To address the scarcity of labeled sentiment data in low-resource Indic languages, we further generated synthetic user reviews in Hindi and Bengali. These reviews were crafted across seven sentiment categories—direct, ironic, sarcastic, incomplete, ambiguous, cross-domain, and mixed—and designed to express multiple emotions such as joy, anger, frustration, admiration, disappointment, satisfaction, ambivalence, and resentment. The prompts explicitly guided the MLLM to generate linguistically nuanced and emotionally complex reviews. Additionally, the models were prompted to estimate the percentage contribution of each emotion present in mixed and complex reviews.

- **Multilingual Sentiment Analysis Evaluation**

  The generated dataset focused on laptop product reviews and served as a benchmark for evaluating multilingual sentiment understanding. We assessed several state-of-the-art MLLMs, including ChatGPT, Llama, Mistral, Phi-3, Gemini, and Gemma. Model predictions were compared against the intended sentiment and emotional composition embedded in the synthetic reviews. To ensure evaluation consistency, the reviews were translated into English and verified through human annotations. The models were instructed to classify emotions with confidence scores (out of 100), returning results in a structured JSON format. This evaluation provided insights into each model's capability to handle nuanced sentiment expressions in low-resource languages.

- **Multilingual Content Generation Using RAG**

  To generate high-quality multilingual product content, we combined web scraping with LLM-based content generation. Product specifications and technical details were collected from multiple online sources and transformed into coherent product titles, features, and descriptions in Hindi and Bengali. The generated content was embedded using sentence transformers and stored in a Chroma vector database.

  We implemented a RAG framework to enable few-shot prompting for new product content. By

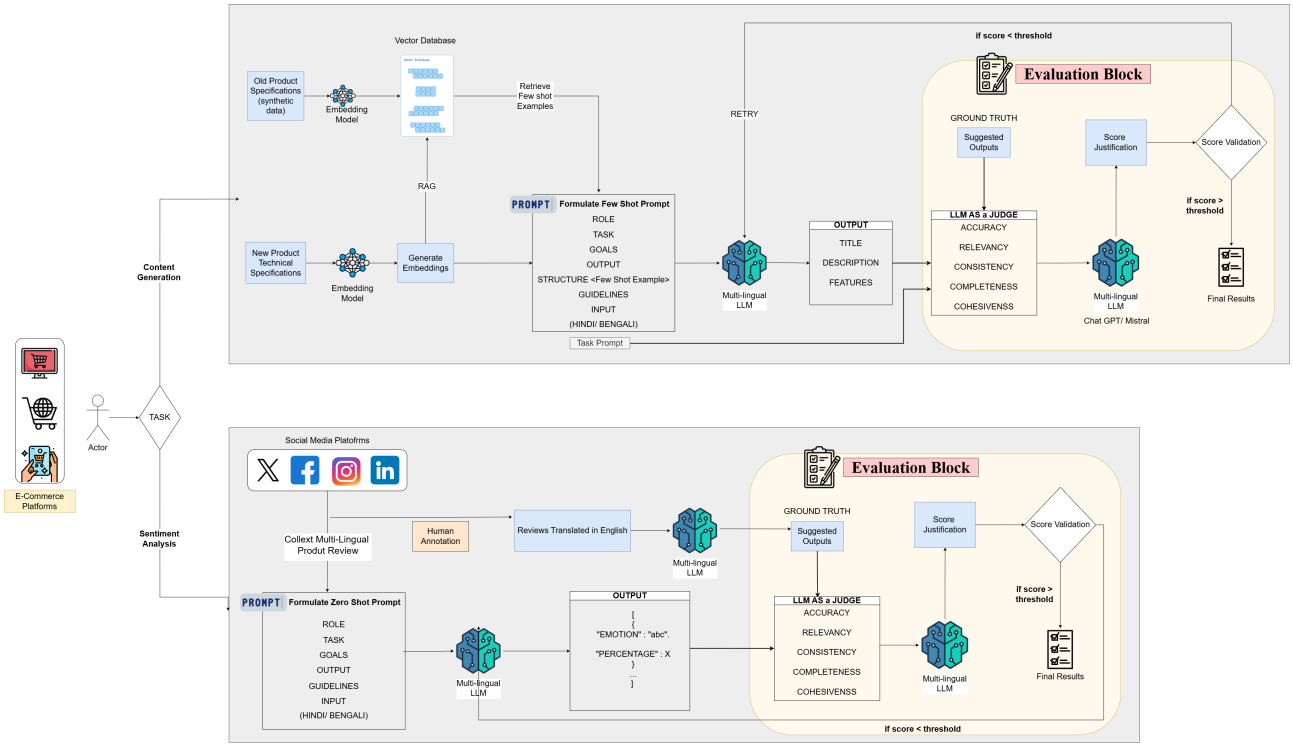

Fig. 1: Architecture Diagram

retrieving semantically similar product embeddings from the vector database, relevant examples were incorporated into the prompts provided to the LLMs. This approach ensured contextual consistency, improved linguistic quality, and enhanced the accuracy of generated multilingual product content, supporting scalable content generation for low-resource Indic languages. These models were evaluated based on their ability to generate user-friendly, native-language content that aligned with the linguistic nuances of Hindi and Bengali using LLM as a Judge technique. The LLM as a Judge evaluated the quality of output across Relevancy, Accuracy, Consistency, Completeness and Cohesiveness.

3. Literature Survey

Unlike prior work that primarily focuses on either multilingual text generation or sentiment classification in high-resource settings, this study presents an integrated, end-to-end GenAI framework for both multilingual e-commerce content generation and fine-grained sentiment analysis in low-resource Indic languages. The novelty lies in (i) the combined use of RAG and COT prompting to produce specification-faithful product content in Hindi and Bengali from minimal inputs, (ii) the systematic evaluation of multiple state-of-the-art Multilingual Large Language Models using an LLM-as-a-Judge framework tailored to low-resource linguistic nuances, and (iii) the modeling of complex, mixed, and ambiguous sentiments with percentage-based emotion distributions, rather than single-label sentiment classification. Additionally, this work introduces a synthetic data generation and validation pipeline specifically designed for low-resource e-commerce scenarios, enabling scalable benchmarking where real annotated data is scarce. Together, these contributions position the work as a practical, deployment-oriented GenAI solution for inclusive multilingual e-commerce platforms.

4. Results

This section presents experimental results for multilingual sentiment analysis and multilingual content generation in low-resource Indic languages (Hindi and Bengali). For sentiment analysis, a comparative evaluation across multiple MLLMs shows that Gemma-2-27B achieves the highest overall accuracy for Bengali reviews, while ChatGPT-4 demonstrates the most consistent performance (>80%) for Hindi reviews. Most models exhibit accuracy clustered in the 60–80% range, highlighting the inherent challenges of complex sentiment classification in low-resource languages.

For multilingual content generation using RAG and CoT, ChatGPT-4o emerges as the best-performing model, consistently generating accurate, coherent, and specification-compliant content in both Hindi and Bengali from a single prompt, as validated by an LLM-as-a-Judge evaluation. Gemini performs competitively but struggles with few-shot adherence and content richness, while Llama shows inconsistency and hallucinations. Mistral and Phi-3.5 MoE Instruct perform poorly, with frequent factual errors and an inability to reliably generate multilingual output.