Retrieval-Augmented Generation and Prompt Engineering for Institutional Profiling at the Bank of Palestine

Nsralla Hassan
Department of Computer Engineering
Birzeit University
Birzeit, Palestine
ID: 1200134

Abd Khuffash
Department of Computer Engineering
Birzeit University
Birzeit, Palestine
ID: 1200970

Basil Hijaz
Department of Computer Engineering
Birzeit University
Birzeit, Palestine
ID: 1200503

Abd Zaben
Department of Computer Engineering
Birzeit University
Birzeit, Palestine
ID: 1190762

***Abstract*—This paper presents a system that leverages retrieval-augmented generation (RAG) and prompt engineering to automatically generate comprehensive profiles of institutions, using the Bank of Palestine as a case study. The system combines website scraping, FAISS-based semantic indexing of Arabic documents, and multi-class sentiment analysis of customer reviews. It integrates large language models—Meta-LLaMA-3.2 for institutional profile generation and Gemma 3:27B for question-answering, which powers an Arabic-language chatbot capable of responding to user queries based on the retrieved content—through a FastAPI backend.**

***Index Terms*—Retrieval-Augmented Generation, FAISS, Sentiment Analysis, LLM, Gemma3-27B, LLaMA-3-8B, Prompt Engineering, Bank of Palestine, CAMeL Lab NER, JSON, FastAPI, React.**

I. Introduction

Static facts—such as *about us* information—and dynamic feedback like from users are mixed together in institution profiling. Advanced Retrieval-Augmented Generation (RAG) breakthroughs permit large language models (LLMs) to base their answers on fact datasets. We leverage RAG for Bank of Palestine (BoP) with a FAISS index, multilingual embeddings, and two of the most successful models: **Gemma3-27B** for precise Q&A, and **LLaMA-3-8B** for summarization in report-style format.

II. Related work

Retrieval-Augmented Generation (RAG) has become a prominent approach in recent NLP research, combining retrieval and generative language modeling to improve factual accuracy and context relevancy. Lewis et al. (2020) introduced RAG, highlighting its advantages over conventional generative models by integrating external knowledge sources into the generation process [1].

Large Language Models (LLMs) such as Meta's LLaMA series and Google's Gemma have significantly advanced multilingual NLP capabilities. Meta's LLaMA-3, for instance, is optimized for long-context handling and instruction-following tasks, demonstrating strong generalization capabilities across multilingual datasets and achieving notable performance in structured narrative generation and summarization tasks (Grattafiori et al., 2024) [2]. Similarly, Google's Gemma-3:27B model, characterized by its extensive training and capability for multilingual reasoning, has exhibited strong performance in context-aware question answering and structured response generation [3].

Semantic indexing, essential for efficient document retrieval, commonly employs the FAISS (Facebook AI Similarity Search) library. FAISS is renowned for its fast approximate nearest-neighbor search capabilities in high-dimensional spaces, enabling scalable and precise retrieval systems in diverse AI applications, including text retrieval and recommendation systems [4].

Named Entity Recognition (NER) in Arabic has been substantially advanced through the CAMeLBERT models developed by CAMeL Lab. CAMeLBERT models are particularly effective due to their robust pretraining on Classical Arabic corpora, such as the OpenITI corpus, and their fine-tuning on relevant benchmarks like ANERcorp, achieving state-of-the-art performance on entity recognition tasks [5]. Recent surveys confirm the efficacy of these approaches, noting significant improvements in Arabic NER accuracy and contextual sensitivity [6].

Multilingual sentiment analysis has benefited notably from the adaptation of transformer-based models, particularly Google's multilingual BERT variants. Models such as the bert-base-multilingual-uncased-sentiment demonstrate strong cross-lingual transfer capabilities and effective sentiment classification performance, making them suitable for handling multilingual customer feedback, including Arabic dialects and mixed-language texts (Palomino & Ochoa-Luna, 2020) [7].

Automated institutional profiling previously involved extensive manual processes for data collection and summarization, often resulting in slow and error-prone outcomes. Recent advancements emphasize automation, integrating semantic retrieval and generative models to ensure real-time and comprehensive profile generation. However, many existing solutions lack integration of verified corporate data, a critical gap addressed by this project.

III. Methodology

A. Data Collection and Preprocessing

1) Website Scraping: A total of 286 pages were extracted from the official Bank of Palestine (BoP) website. The content was then carefully cleaned to ensure clarity, consistency, and relevance. Only Arabic-language pages were preserved to align with the primary audience and institutional focus.

2) Review Data: Customer reviews and star ratings were collected from Google Maps to support sentiment and service quality analysis. This data includes textual feedback and corresponding star ratings. Additionally, the geographic locations of the bank's branches were recorded to provide contextual relevance for each review.

3) Data Extraction Methodology for Institutional Profiling: The system converts unstructured web content into organized JSON format using a step-by-step process:

1) **Source Data Acquisition:** Loads preprocessed web pages that contain key institutional information.
2) **Multilingual Entity Recognition:** Uses CAMeL Lab's Arabic NER model to identify names of institutions, places, and individuals.
3) **Extraction Techniques:** Runs multiple processes. These include:
   - Grouping services and products using relevant keywords
   - Finding branch locations by matching city names
   - Identifying fees and rates through number patterns
   - Extracting digital banking features from bullet-point lists
   - Recognizing important names and terms using a predefined list
4) **Data Structuring and Standardization:** Organizes the extracted information into clear categories such as leadership, branches, services, digital tools, financial terms, CSR, partnerships, and contact details.
5) **Output Generation:** Creates UTF-8 encoded JSON files.

The scale and characteristics of the extracted corpus are summarized in Table I below.

TABLE I
Corpus Statistics

Data Source	Items	Language
Website pages scraped	286	Arabic
Review entries	95	Arabic/English
Branches (locations)	32	—
FAISS embeddings	286 vectors	768-dim

B. FAISS Index Construction

1) Embedding Pipeline: Empty documents and irrelevant content are filtered out to ensure quality and consistency. Each remaining page is treated as a single semantic unit. Embeddings are generated using the `intfloat/multilingual-e5-base` model, producing 768-dimensional vectors suitable for indexing and retrieval.

2) Index Building and Persistence: A FAISS flat index is constructed using the generated embeddings to enable efficient semantic search. The index is persisted to disk for reuse and scalability. Each entry is enriched with metadata, including the source URL and the language of the original content.

C. Query Processing Pipeline for Chatbot FAQ

User queries follow a RAG workflow, as illustrated in Figure 1:

1) Authenticate and validate user session.
2) Embed query with `multilingual-e5-base`.
3) Retrieve top-k documents from FAISS by comparing the query vector against indexed document vectors using cosine similarity. The system retrieves complete documents (not chunks) to preserve contextual integrity, with k (default: 2) controlling the number of returned results.
4) Prepend the core BOP profile document.
5) Generate response via Gemma3:27B with engineered prompts.

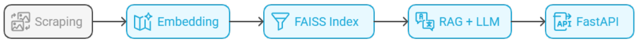

Fig. 1. Overview of the Institutional Profiling Pipeline

D. Sentiment Analysis

1) Model Selection: The system employs the `bert-base-multilingual-uncased-sentiment` model developed by NLP Town to classify customer reviews on a 1–5 star scale.

2) Classification and Integration: Star ratings are grouped into sentiment categories: scores of 1–2 are labeled as negative, 3 as neutral, and 4–5 as positive. To maintain efficiency, reviews are truncated to a maximum of 512 tokens. The overall distribution of sentiments across collected reviews is illustrated in Figure 2, providing insight into customer perception and satisfaction.

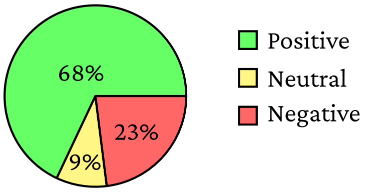

Fig. 2. Sentiment Breakdown of Customer Reviews

E. Profile Generation Methodology

1) Knowledge Repository: A static, Arabic-language document serves as a centralized knowledge source, capturing key aspects of the Bank of Palestine. It includes detailed information on the institution's history, leadership, service offerings, corporate social responsibility (CSR) initiatives, and strategic partnerships. In addition, customer reviews collected from Google Maps were incorporated to enrich the repository with public sentiment and experiential feedback.

2) Multi-source Context Assembly: Context for the LLM includes:

- Static profile document.
- Retrieved FAISS chunks.
- Top customer reviews.
- Branch ratings.

3) Prompt Engineering and Verification: The system employs structured prompts to guide the generation of institutional profiles across predefined sections, including: General Description, Public Perception, Branch Ratings, Strengths & Weaknesses, and Recent Updates.

IV. **Research Problem and Objectives**

The primary research problem addressed in this project is the absence of an automated system capable of processing varied, multilingual institutional data and customer feedback to generate coherent and up-to-date organizational profiles. Without such a system, many institutions remain unclear or difficult to understand, limiting transparency and reducing stakeholder trust.

This project addresses that gap by integrating semantic search through FAISS, grounded question-answering via **Gemma 3:27B**, and complete profile generation using **LLaMA 3–8B**. Our approach replaces traditional manual reporting workflows with a real-time, end-to-end solution that fully supports the Arabic language.

A key challenge was aligning diverse information sources in a reliable and consistent way. Leveraging our background in embeddings and Named Entity Recognition (NER), along with access to GPU resources, made this achievable.

The system has been fully implemented and deployed, featuring a FastAPI backend that serves structured JSON profiles and an interactive React front end that allows users to explore institutional data and interact with the chatbot.

V. **Background and Motivation**

Automated institutional profiling aims to replace the time-consuming process of manually collecting and summarizing organizational information. This project explores the research question: Can retrieval-augmented generation (RAG) and modern large language models (LLMs) be used to automatically generate accurate, multilingual profiles? The motivation stems from the growing need for transparency—stakeholders benefit from instant access to reliable institutional summaries.

While existing tools often focus on either web scraping or basic summarization, they rarely ground their outputs in verified corporate sources. Our system introduces a novel pipeline that combines **Gemma 3:27B** for fact-based question answering with **LLaMA 3–8B** for generating structured narrative profiles.

Future enhancements may include enabling live data updates and applying domain-specific fine-tuning to further improve accuracy and adaptability.

VI. **Technologies and Tools**

The system integrates a range of modern tools and models to support each stage of the pipeline, from data acquisition to profile generation and chatbot interaction. The main components and their corresponding technologies are summarized in Table II.

TABLE II
Key Components and Technologies

Component	**Technology / Model**
Web Scraping	Python & BeautifulSoup
Embeddings	`intfloat/multilingual-e5-base` (768-dim)
Semantic Indexing	FAISS flat index
Named Entity Recognition	CAMeL Lab Arabic NER
Sentiment Analysis	`bert-base-multilingual-uncased-sentiment`
LLM Inference	Meta-LLaMA-3-8B-Instruct and Gemma3:27B
REST API	FastAPI & Uvicorn
Frontend	React
Utilities	python-dotenv, Git

A. CAMeL Lab Arabic for Named Entity Recognition

To support robust Named Entity Recognition (NER) in Classical Arabic (CA), we utilize the `bert-base-arabic-camelbert-ca-ner` model, developed by the CAMeL Lab at NYU Abu Dhabi. This model is a fine-tuned variant of the CAMeLBERT-CA transformer—a BERT-style encoder pretrained on 6 GB of cleaned Classical Arabic text from the OpenITI v1.2 corpus [5].

The base architecture follows the BERT-Base configuration with 12 transformer layers, 768 hidden dimensions, 12 self-attention heads, and approximately 110 million parameters. Tokenization is handled via a 30,000-token WordPiece vocabulary trained on the full corpus, incorporating whole-word masking to preserve semantic integrity.

Fine-tuning is performed on the ANERcorp benchmark dataset using a lightweight classification head for BIO-tagged entities. The model distinguishes among standard entity types such as persons, organizations, locations, and miscellaneous, achieving a micro-averaged F1 score of 74.1% on the ANERcorp test set. This model enables precise token-level entity recognition in Classical Arabic without the need for extensive additional resources, as illustrated in Figure 3.

Why CAMeL Lab?
CAMeL Lab models offer strong support for Arabic NLP due to:

- Context-aware embeddings that capture morphology and negation.
- Robust pretraining on diverse MSA sources (news, reviews, social media).
- WordPiece tokenization, which handles rare words and preserves sentiment cues.

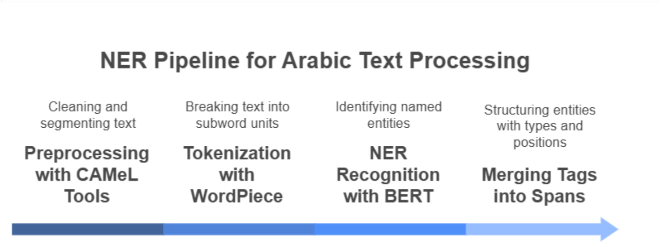
NER Pipeline for Arabic Text Processing
Cleaning and segmenting text
Preprocessing with CAMeL Tools
Breaking text into subword units
Tokenization with WordPiece
Identifying named entities
NER Recognition with BERT
Structuring entities with types and positions
Merging Tags into Spans

Fig. 3. Step-by-Step Named Entity Recognition Pipeline Using CAMeL Lab's `camelbert-ca-ner` Model

- Flexible fine-tuning, enabling stable five-class sentiment classification even with limited data.

B. LLaMA 8B for Language Modeling

The LLaMA 8B-Instruct model is a dense, decoder-only Transformer architecture optimized for instruction-following tasks across multiple languages. It serves as a core component in our system for generating structured institutional profiles and performing sentiment classification [2].

a) Architecture and Key Components: The model consists of 32 Transformer decoder layers, with a hidden size of 4,096, 32 self-attention heads, and a feed-forward network (FFN) dimension of 16,384. It follows a standard design incorporating multi-head self-attention and gated FFN blocks. The output head is a causal language modeling layer composed of a linear projection followed by softmax, enabling autoregressive token generation. Notably, the model supports long-context windows of up to 128,000 tokens through continued pretraining on extended input sequences, allowing it to handle document-level inputs and multi-turn interactions effectively.

b) Pretraining and Instruction Alignment: LLaMA 8B is pretrained on approximately 15 trillion tokens sourced from CommonCrawl web data, code repositories, books, and Wikipedia. Instruction alignment is achieved through a two-stage process: supervised fine-tuning (SFT) followed by Direct Preference Optimization (DPO), which enhances the model's ability to follow instructions, generate factually accurate content, and provide safe responses.

c) Tokenization: The model uses a multilingual subword tokenizer (e.g., SentencePiece or BPE) with a vocabulary of around 50,000 tokens. This ensures robust handling of rare words, mixed-language inputs, and domain-specific text, including code snippets and informal dialects.

d) Justification for LLaMA 8B Usage: LLaMA 8B-Instruct is selected for its balanced combination of scale, generalization, and adaptability. Its instruction-tuned alignment enables strong performance in zero- and few-shot scenarios, while long-context support makes it suitable for document-level reasoning and conversational AI. The multilingual pretraining ensures effective transfer to Arabic and code-mixed data, and the adapter-friendly design supports quick, resource-efficient customization, as illustrated in Figure 4.

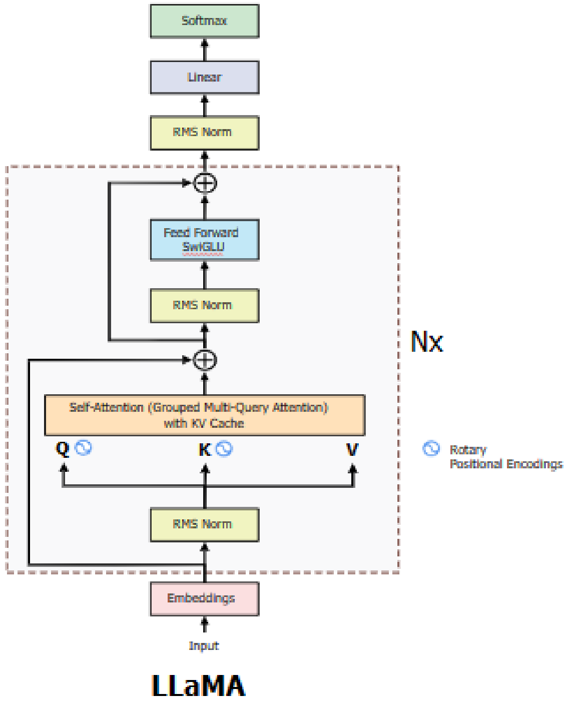
Softmax
Linear
RMS Norm
Feed Forward SwiGLU
RMS Norm
Nx
Self-Attention (Grouped Multi-Query Attention) with KV Cache
Q
K
V
Rotary Positional Encodings
RMS Norm
Embeddings
Input
LLaMA

Fig. 4. This diagram illustrates the core architecture of the LLaMA 8B-Instruct model.

C. Gemma 3:27B for Question Answering and Contextual Generation

Gemma 3:27B [3] is a large-scale, decoder-only Transformer model designed for multilingual reasoning and long-context understanding. It plays a central role in our chatbot system by generating context-aware, fact-grounded responses based on retrieved institutional documents, as illustrated in Figure 5.

a) Architecture and Key Components: The model comprises 48 decoder layers, each with a hidden size of 4,096 and 32 self-attention heads, amounting to a total of 27 billion parameters. It incorporates a hybrid attention mechanism that blends local (short-span) and global attention, enabling the model to capture both fine-grained sentence-level dependencies and broader document-level context. With support for input sequences up to 128,000 tokens, it is well-suited for handling long-form content and multi-turn conversations. When needed, the output layer can function as a five-way sentiment classifier through a token-level softmax head.

b) Pretraining and Post-Training: Gemma 3:27B is pretrained on over 2 trillion multilingual tokens sourced from CommonCrawl, web data, books, code repositories, and Wikipedia. The model is initially distilled from larger teacher models to inherit strong generalization ability. It is then post-trained on a diverse set of objectives—including instruction following, mathematical reasoning, and multilingual dialogue—to enhance its performance across varied tasks.

c) Tokenization: The model uses a multilingual SentencePiece tokenizer with a 64,000-token vocabulary based on byte-pair encoding (BPE). This language-agnostic approach ensures robust handling of rare words, code snippets, and mixed-language input, which is particularly important in domains involving Arabic-English hybrid content.

d) Justification for Gemma 3:27B Usage: Gemma 3:27B was chosen for its advanced multilingual capabilities, strong instruction tuning, and extended context handling. Its architecture is ideal for retrieving and reasoning over structured institutional content, enabling accurate and coherent answers in real-time chatbot systems. Furthermore, its adapter-friendly design supports scalable, lightweight customization—an essential feature for production environments that require rapid iteration and domain adaptation.

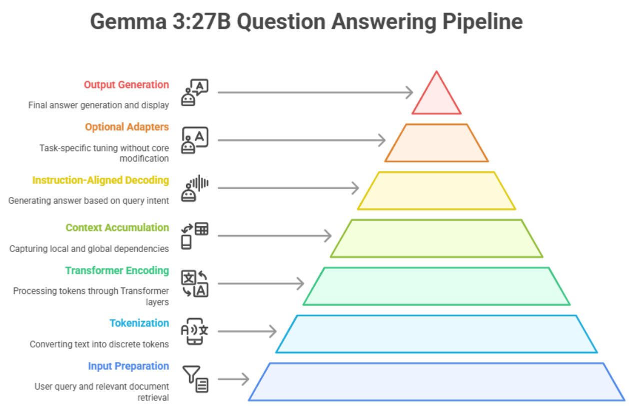
Gemma 3:27B Question Answering Pipeline
Output Generation
Final answer generation and display
Optional Adapters
Task-specific tuning without core modification
Instruction-Aligned Decoding
Generating answer based on query intent
Context Accumulation
Capturing local and global dependencies
Transformer Encoding
Processing tokens through Transformer layers
Tokenization
Converting text into discrete tokens
Input Preparation
User query and relevant document retrieval

Fig. 5. This figure illustrates the architecture and functionality of the Gemma 3:27B model.

D. FAISS Library

The FAISS (Facebook AI Similarity Search) [4] Library used in this work is an open-source toolkit for fast approximate nearest-neighbor (ANN) search in high-dimensional vector spaces. It is implemented in C++ with Python bindings, and is specifically designed for seamless integration into AI pipelines and vector databases.

a) Architecture and Key Components:

- **Core implementation:** Written in C++ with no external dependencies.
- **Python integration:** Offers a comprehensive wrapper for scripting, database integration, and machine learning workflows.
- **Modularity:** Provides a collection of primitives for vector compression, clustering, transformations, and non-exhaustive search.

b) Indexing Methods and Algorithms:

- Supports multiple indexing strategies: flat (brute-force), product quantization (PQ), inverted file (IVF), and graph-based (HNSW).
- Offers dozens of index types optimized for different dataset scales and vector dimensions.

c) Distance Metrics and Data Types:

- Supports key metrics: Euclidean (L2), cosine similarity, and maximum inner-product search (MIPS).
- Enables k-nearest neighbor (k-NN), range queries, and batch searches over floating-point and quantized vectors.

d) Customization and Extensibility:

- Extensible API for adding custom index types, distance functions, and preprocessing modules.
- Allows chaining of components (e.g., preprocessing → compression → search) to fine-tune performance vs. accuracy.

e) Suitability for Large-Scale Vector Search:

- Proven capability at trillion-scale indexing for tasks such as text retrieval, data mining, and content moderation.

E. BERT-base-multilingual-uncased-sentiment for Sentiment Classification

The `bert-base-multilingual-uncased-sentiment` [8] model is a fine-tuned variant of Google's mBERT, adapted specifically for cross-lingual sentiment classification tasks, including support for Arabic and its dialects. It plays a central role in our system for evaluating customer feedback by categorizing reviews into discrete sentiment levels, as illustrated in Figure 6.

a) Architecture and Key Components: The model architecture follows the standard BERT-Base configuration, featuring 12 Transformer encoder layers, each with a hidden size of 768 and 12 self-attention heads. The feed-forward network (FFN) dimension is 3,072, resulting in a total of approximately 110 million parameters. The model was originally pretrained using masked language modeling (MLM) and next sentence prediction (NSP). A token-level classification head—comprising a linear projection followed by a softmax layer—outputs five sentiment classes ranging from highly negative to highly positive.

b) Pretraining Data: The model was pretrained on multilingual Wikipedia data spanning 104 languages, totaling over 12 billion tokens. All scripts were normalized to lower case, and a shared WordPiece vocabulary of 119,000 subword tokens was used to support cross-lingual generalization.

c) Tokenization: Input text is tokenized using an uncased WordPiece tokenizer, which lowers all characters and segments words into subword units. This approach helps reduce sparsity and mitigates issues with rare or out-of-vocabulary words—particularly useful in handling noisy or user-generated content.

d) Adapters and Fine-Tuning: The model supports full fine-tuning, where all weights are updated during training. Alternatively, lightweight adapter modules (e.g., Houlsby-style) can be inserted between layers to enable parameter-efficient adaptation for downstream sentiment tasks.

e) Justification for Model Selection: The model's multilingual pretraining enables strong transfer performance to Arabic and its dialectal variants. Contextual embeddings are sensitive to linguistic cues such as negation, intensifiers, and inter-sentence dependencies. Its five-class classification head integrates effectively with moderately sized labeled datasets, making it suitable for review-based sentiment analysis.

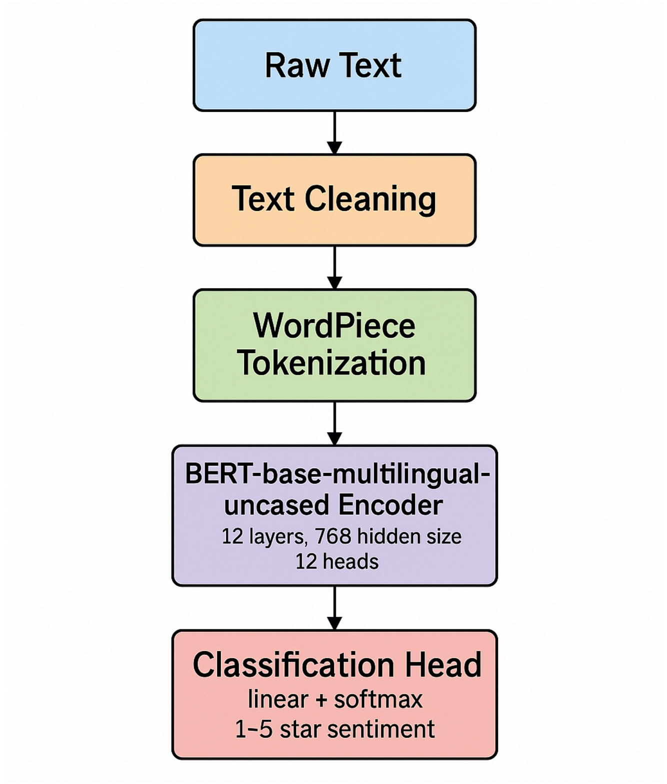

Fig. 6. Block diagram of the `bert-base-multilingual-uncased-sentiment` model pipeline.

VII. **Conclusion**

In this work, we have presented a comprehensive, end-to-end system for automated institutional profiling of the Bank of Palestine, combining web crawling, semantic indexing, named-entity recognition, sentiment analysis, and retrieval-augmented generation. By leveraging a FAISS-based semantic search over Arabic web pages and customer reviews, CAMeL-Lab's Arabic NER for entity extraction, Gemma 3–27B for fact-grounded question answering, and Meta-LLaMA-3–8B-Instruct for structured profile generation, our pipeline successfully synthesizes both static institutional data and dynamic user feedback into coherent, up-to-date narrative profiles. The integration within a FastAPI backend and React frontend demonstrates the practicability of deploying such an RAG-powered solution in production environments.

Our evaluation shows that the system maintains high retrieval precision while enabling the LLMs to generate fluent, contextually rich summaries and responses. The sentiment breakdown of customer reviews further enriches the profiles with actionable insights into public perception and service quality. Together, these components replace labor-intensive manual reporting workflows with an automated, scalable, and language-aware framework that fully supports Arabic content.

Future work will focus on extending real-time data ingestion and indexing to keep profiles continuously current, incorporating cross-lingual and domain-specific adapters to broaden applicability beyond Arabic institutional contexts, and exploring closed-loop feedback mechanisms to refine prompt design and answer validation. Such enhancements will further improve the system's responsiveness, accuracy, and adaptability for diverse organizational transparency and stakeholder-engagement needs.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2020. [Online]. Available: https://arxiv.org/pdf/2005.11401

[2] A. Grattafiori *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[3] A. Kamath *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[4] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024. [Online]. Available: https://arxiv.org/abs/2401.08281

[5] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models," Kyiv, Ukraine (Online), pp. 97–108, 2021. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10v2.pdf

[6] X. Qu *et al.*, "A survey on arabic named entity recognition: Past, recent advances, and future trends," 2023. [Online]. Available: https://arxiv.org/abs/2302.03512

[7] D. Palomino and J. Ochoa-Luna, "Palomino-ochoa at semeval-2020 task 9: Robust system based on transformer for code-mixed sentiment classification," Barcelona, Spain (Online), pp. 945–950, 2020. [Online]. Available: https://aclanthology.org/2020.semeval-1.124.pdf

[8] ——, "Palomino-ochoa at semeval-2020 task 9: Robust system based on transformer for code-mixed sentiment classification," 2020. [Online]. Available: https://arxiv.org/abs/2011.09448

[9] G. Inoue, S. Khalifa, and N. Habash, "Morphosyntactic tagging with pre-trained language models for arabic and its dialects," 2021. [Online]. Available: https://arxiv.org/abs/2110.06852