Restaurant Recommendation Chatbot Using NLP-Based Review Analysis

Burak Metin Çapraz, Hülya Hasnalbant, Serra Akyıldız

Abstract

Online restaurant reviews represent a valuable but underutilized source of feedback for both business owners and customers. However, reviews typically appear as bulk unstructured text, requiring significant effort to extract useful insights. While platforms often request a simple star rating (1–5), these ratings rarely capture the nuances expressed in written feedback.
This project introduces a ***feedback chatbot****: a B2C chatbot that recommends restaurants based on user queries and old reviews. Our methodology integrates Natural Language Processing (NLP), sentiment analysis, aspect-based classification, embeddings, and knowledge graph construction.*
We scraped a total of ***2420 restaurants*** *across* ***İzmir, İstanbul, and Ankara****, collecting approximately* ***153,486 reviews****. The restaurants represent diverse categories. Using this dataset, we used models to detect sentiment polarity, categorize reviews into aspects (taste, service, ambiance, price-performance, hygiene, and menu variety), and generate restaurant-level summaries. A Neo4j knowledge graph was built to capture relationships between restaurants, aspects, menu items, and locations. Finally, we designed a chatbot using embedding similarity and transformers to provide personalized restaurant recommendations. We compared both knowledge graph and embedding technique chatbot results.*

Keywords

NLP, Sentiment Analysis, Aspect-Based Sentiment, Embeddings, Transformers, Restaurant Reviews, Knowledge Graph, Conversational Agents

1. Introduction

Customer reviews significantly influence restaurant reputation and consumer decision-making. In the digital era, platforms such as Google Maps provide vast amounts of user-generated data. However, most of this information remains unstructured, limiting its utility. Traditional sentiment analysis approaches often stop at classifying reviews as positive or negative, offering little actionable guidance (Liu, 2012).

Our project addresses this gap by introducing a smart feedback agent that transforms restaurant reviews into structured insights and delivers personalized restaurant recommendations for B2C users. This work is particularly important for Turkish-language datasets, where diacritics and morphological richness pose unique challenges for NLP pipelines. This project makes three contributions:

1. Development of a comprehensive pipeline integrating scraping, cleaning, sentiment analysis, and embeddings.
2. Construction of a Neo4j-based knowledge graph to capture entity relationships (Wang et al., 2019).

3. Deployment of a chatbot that leverages embeddings to recommend restaurants, bridging the gap between data insights and end-user interaction.

2. Related Work

Early research in sentiment analysis relied on classical approaches such as bag-of-words models with Support Vector Machines (SVM) and Logistic Regression (Pang et al., 2002). With the emergence of deep learning, recurrent neural networks and LSTMs improved sequential modeling for sentiment tasks (Hossain et al., 2020). More recently, transformer architectures such as BERT (Devlin et al., 2019) have set a new state-of-the-art for text embeddings and classification, enabling more accurate sentiment detection and contextual understanding.

Aspect-Based Sentiment Analysis (ABSA) has provided a more fine-grained approach by linking sentiments to attributes such as taste, service, and ambiance. Hua et al. (2024) review ABSA techniques and highlight their increasing role in recommendation systems and business intelligence. In the restaurant domain, Cuizon et al. (2019) showed that aspect-linked ratings offer more actionable insights than overall sentiment scores alone. Recent advancements in large language models (LLMs) have further expanded ABSA research. Instruction-tuned LLMs, particularly those incorporating retrieval-based example ranking, have demonstrated strong performance by aligning model outputs with task-specific prompts and semantically relevant examples, enabling flexible adaptation to diverse ABSA subtasks with minimal supervision (Rahman et al., 2025).

Knowledge graphs have also become important in recommender systems, as they allow structured representation of entities and their relationships. Wang et al. (2019) introduced KGAT, an attention-based framework that leverages graph structures for improved recommendations. Similarly, Loesch et al. (2022) demonstrated how knowledge graph embeddings can identify food substitutions and alternatives in culinary contexts.

Our work builds on these developments by combining ABSA, embeddings, and knowledge graphs into a unified pipeline. Unlike prior studies, we extend the approach with a practical chatbot interface that enables real-time, personalized restaurant recommendations for end users.

3. Proposed Solution

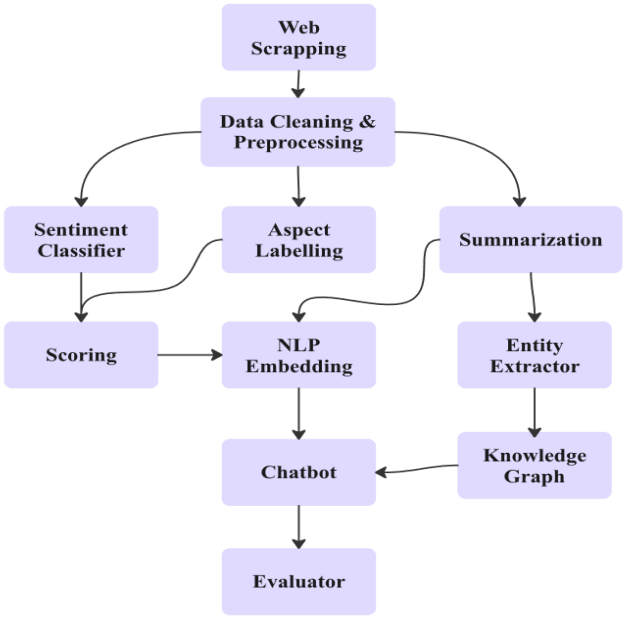

Fig. 1. Proposed Solution Architecture

4. Methodology

4.1 Scraping

Data was collected from Google Maps using the Selenium library, chosen for its ability to handle dynamic HTML structures and automate browser interactions. The study focused on Turkey's three largest cities—Istanbul, Ankara, and Izmir—due to their population density and restaurant diversity. For each restaurant, up to 100 of the most recent reviews were collected, ensuring both currency and balance across establishments. Reviews suspected to be written by restaurant owners or staff were excluded. In total **153,486** reviews from **2,420** restaurants were gathered for analysis.

4.2 Data Cleaning

In order to ensure the reliability of the dataset and the validity of the analysis results, a comprehensive data cleaning process was carried out. Within this scope:

- **Duplicate Reviews:** Repeated reviews made by the same customer for the same restaurant were removed, and only one instance was retained.
- **Outdated Reviews:** Reviews dated 2018 and earlier were excluded, as they were considered outdated and likely to lead to misleading insights. This group accounted for approximately 1% of the total reviews.
- **Restaurants with Low Review Counts:** Records belonging to restaurants with fewer than 13 total reviews were excluded from the analysis due to potential bias risk. These

removals corresponded to about 1% of the dataset.

- **Short Reviews:** Reviews consisting of only one or two words, which generally lacked meaningful content, were filtered out.

4.3 Sentiment Classification

We conducted experiments with several Turkish BERT models for sentiment classification, including Savasy Turkish Sentiment, DBMDZ Turkish Base, DBMDZ Turkish Uncased, YTU Turkish BERT, and Lodos Turkish Sentiment. While some models encountered compatibility issues, the Savasy Turkish Sentiment model **(savasy/bert-base-turkish-sentiment-cased)** demonstrated the highest performance, achieving robust confidence intervals across all evaluation metrics. Based on these results, we selected this model as our primary classifier. It was applied to classify Turkish restaurant reviews into positive, negative, or neutral categories, with both sentiment labels and confidence-weighted scores stored in the dataset for further analysis. As shown in Figure 2, the classification results revealed that 58.1% of the reviews were positive while 41.9% were negative, indicating that the majority of customers expressed satisfaction but a significant portion still reported negative experiences.

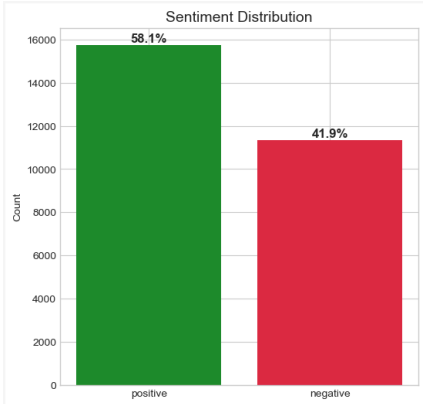

Fig. 2. Distribution of positive and negative reviews

4.4 Aspect Labeling

Aspect-based sentiment analysis was performed using **LLM prompting (Gemini 2.5 Flash)**. Reviews were categorized into taste, service, ambiance, price-performance, hygiene, and menu variety. Aspect-level sentiment scores were aggregated to produce restaurant-level insights. As illustrated in Figure 3, ambiance and taste received the highest average sentiment scores, while hygiene and price-performance were rated the lowest. In terms of review frequency, taste and service dominated the dataset with the largest number of reviews, whereas hygiene and menu variety were mentioned far less often.

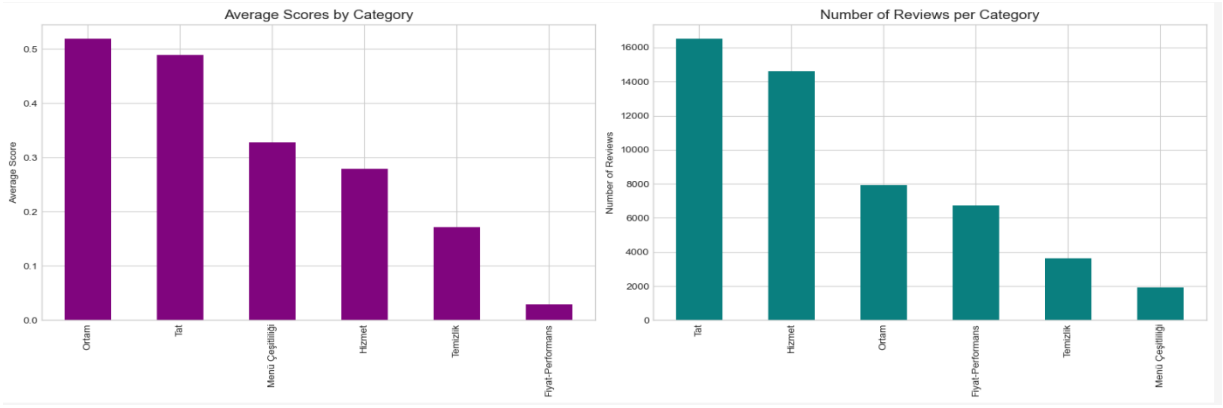

Fig.3. Distribution of Aspects

4.5 Summarization

Customer reviews frequently contained spelling errors, inconsistent phrasing, and semantic ambiguities, making direct use problematic. To address this, all reviews for each restaurant were aggregated and transformed into coherent, representative summaries. For the summarization task, several Turkish fine-tuned large language models (LLMs) were evaluated, including **Turkish Gemma 9B**, **LLaMA3 8B Instruct (Turkish Finetuned)**, and **OpenAI's GPT-OSS 20B**. Model outputs were manually assessed against three criteria: **linguistic accuracy**, **semantic fidelity**, and **readability**. Among the candidates, **Turkish Gemma 9B** consistently outperformed alternatives, particularly in merging fragmented comments, minimizing redundancy, and preserving sentiment nuances. Consequently, it was selected as the primary model for generating standardized review summaries. These summaries provide coherent textual representations that enhance downstream processes such as sentiment analysis, thematic clustering, and personalized recommendation.

Example Summary:

"*Restoran, özellikle Adana dürüm ve kebaplarıyla öne çıkıyor. Müşteriler lezzetli ve uygun fiyatlı yemekleri övgüyle anlatıyorlar. Çalışanların güler yüzlü ve samimi davranışları da sıkça övülüyor. Mekanın temizliği konusunda ise bazı eleştiriler mevcut, ancak hijyen konusuna daha fazla önem verilmesi durumunda restoranın deneyimini daha da olumlu hale getirebileceği düşünülüyor. Özetle, lezzetli yemekleri, uygun fiyatları ve samimi hizmetiyle Etimesgut bölgesinde tercih edilen bir mekan olarak öne çıkıyor.*"

4.5.1 Evaluation of Summaries

The quality of generated summaries was further validated using **BERT-based semantic similarity analysis**. Specifically, cosine similarity was calculated between original reviews and their summaries using **dbmdz/bert-base-turkish-cased**. Results indicate strong semantic preservation:

- 2,367 restaurants (98.7%) achieved *very high* similarity (0.8–1.0).
- 53 restaurants (2.2%) achieved *high* similarity (0.6–0.8).
- No restaurants scored below 0.6.

The overall mean similarity score of 0.91 demonstrates that the summaries retain approximately 91% semantic fidelity to the original reviews. This high level of consistency confirms the reliability of the summarization pipeline and provides a solid foundation for subsequent tasks such as aspect-based sentiment analysis, entity extraction, and knowledge graph construction.

4.6 Weighted Scoring

After cleaning the dataset by removing null values and correcting data type issues, we normalized all numerical aspect scores (taste, service, atmosphere, price-performance, menu variety, and cleanliness) using min-max scaling. To build the ranking, we designed a weighted scoring formula: "***Weighted_Score = Score × Comment_Count × Category_Ratio"***. The category ratio was defined as each category's non-zero comment count divided by the total across all categories, ensuring categories with more feedback had stronger influence. Summing the six weighted scores produced the ***Total_Weighted_Score***, a balanced metric that reflects both review quality and volume for comprehensive restaurant comparisons. As shown in Figure 4, most restaurants scored between 0.1 and 0.2, with very few reaching the higher end of the distribution.

Fig.4. Distribution of Normalized Overall Weighted Aspect Scores

4.7 Entity Extraction

We used Gemini 2.5 Flash with LLM prompting to extract menu items, amenities, and contextual terms in json format. Regex rules and curated dictionaries supported entity detection, with common examples including kebap, pizza, and child-friendly.

4.8 Knowledge Graph

Entities and relationships were imported into **Neo4j**, creating a structured graph of the restaurant domain. Relationships included RATED_FOR, SERVES, LOCATED_IN, and HAS_AMENITY. The graph enabled structured queries and advanced analysis.
Using Neo4j, we transformed user input into Cypher queries for chatbot integration, allowing real-time interaction and personalized responses based on the graph data.
The extraction of thousands of menu items and amenities, including common terms was validated through Neo4j.

The Knowledge Graph contained a total of 13,386 nodes and 50,183 relationships. Key statistics include:

Relationship Type	Count	Percentage
RATED_FOR	37284	%74,3
SERVES	3016	%6,01
CO_OCCUR	2726	%5,43
LOCATED_IN	5932	%11,82
SUITS_OCCASION	627	%1,25
HAS_AMENITY	443	%0,88
CONTRASTS	151	%0,3
IMPLIES	4	%0,01

Node Type	Count	Percentage
Aspect	10194	%76,15
Restaurant	2339	%17,47
MenuItem	757	%5,66
Location	84	%0,63
Amenity	7	%0,05
Occasion	5	%0,04

Fig.5. Distribution of Relationship and Node Type

The graph enabled advanced querying, such as identifying restaurants that serve specific dishes and excel in ambiance.

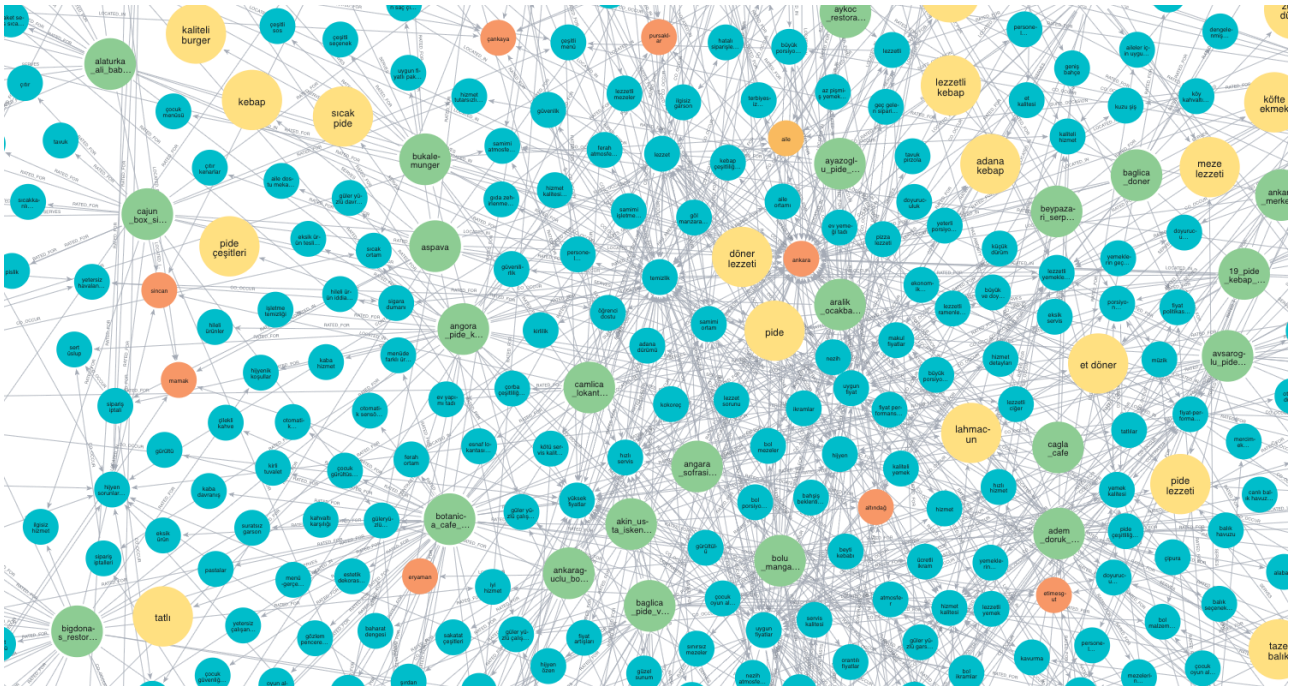

Fig.6. Knowledge graph screenshot example

4.9 NLP Embedding & Similarity

The proposed system is designed as a multi-stage pipeline integrating preprocessing, error-tolerant location correction, and semantically informed recommendation to ensure both robustness and user relevance.

In the preprocessing stage, the dataset is loaded and diacritical inconsistencies in Turkish characters (e.g., “ş/s”, “ı/i”, “ç/c”) are normalized through a custom function. This step is crucial for reliable geographical comparisons, enabling accurate extraction of unique provinces and districts.
The second stage addresses the challenge of noisy user input. Location queries are automatically validated using a hybrid correction mechanism: (i) a local large language model, Turkish Gemma 9B, detects and amends spelling errors; (ii) fuzzy string matching serves as a fallback when the LLM is bypassed or uncertain. This dual strategy ensures resilience to incomplete or erroneous entries while preserving alignment with valid dataset entities.

The third stage applies advanced natural language processing. Restaurant reviews are summarized and keywords extracted, then embedded into a vector space with Sentence-BERT (**all-MiniLM-L6-v2**). User queries are normalized via the **Zemberek** library and embedded with the same model, allowing semantic similarity to be computed. A weighted scoring function (80% summaries, 20% keywords) balances contextual depth with keyword specificity.

Finally, a ranking mechanism retrieves the top-5 most semantically relevant restaurants and re-orders them by Total Weighted Score. This guarantees recommendations that not only capture semantic closeness but also incorporate aggregated quality signals. The system thus provides high-fidelity, semantically meaningful, and quality-aware restaurant recommendations, demonstrating the effectiveness of combining LLM-based correction, fuzzy matching, and modern embedding techniques.

4.10 Chatbot

On top of the embedding-based infrastructure, a web interface and chatbot were developed to enable interactive, conversational recommendations rather than static menu selections. The chatbot collects necessary information from users, processes their responses, and runs the embedding framework to deliver personalized restaurant suggestions. The interface was built with CSS, HTML, and JavaScript, and given limited prior expertise, LLMs were used to assist in code generation, ensuring the design remained both functional and easy to maintain. The overall structure was deliberately kept simple to enhance usability and accessibility while effectively linking natural language input with data-driven recommendation logic.

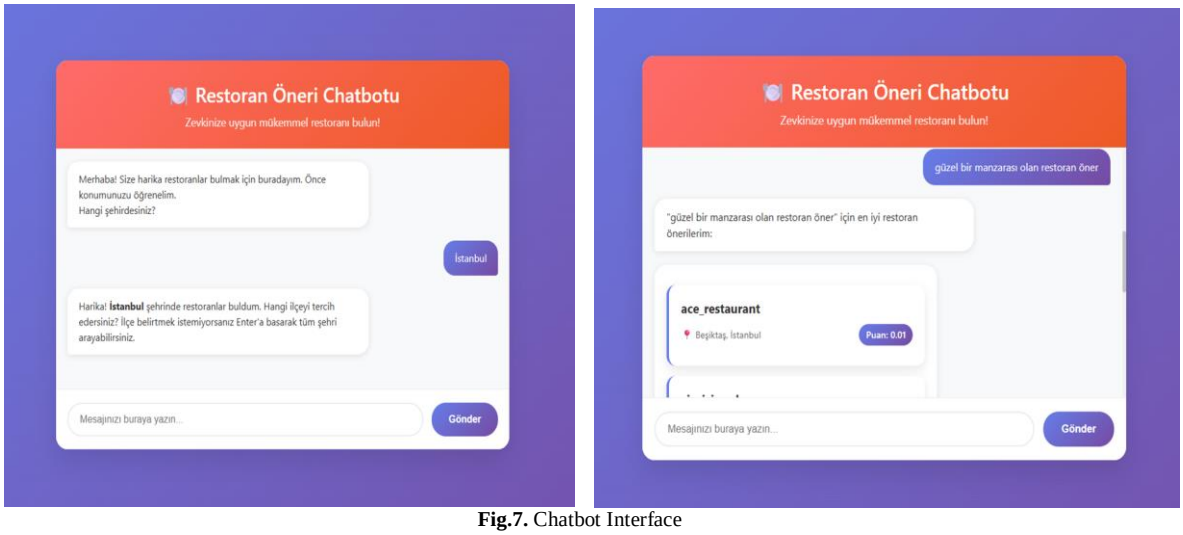

Fig.7. Chatbot Interface

4.11 Rubric-Based ML Evaluation

The evaluation process was guided by four rubric dimensions—Relevance, Factual Accuracy, Diversity, and Coherence. Relevance measured whether the recommendation matched the query (e.g., a user seeking a kebab place received a kebab suggestion). Factual accuracy checked if the recommended restaurant actually existed in the requested location and met the query's specifications. Diversity ensured a range of relevant options, while coherence assessed internal consistency and detail (e.g., whether a restaurant was correctly described as clean and quiet).

To operationalize this rubric, a subset of 20% from 100 total queries was manually rated on a 1–5 scale across both the KG-based and embedding-based models. These ratings were used as training data: text was encoded using TF-IDF, and the remaining queries were labeled via an XGBoost classifier with hyperparameter tuning. Average rubric scores were then combined using weighted averages—Factual Accuracy (0.5), Relevance (0.3), Diversity (0.2), and Coherence (0.2). Factual accuracy carried the highest weight to guarantee trustworthy results, followed by relevance for intent alignment, with diversity and coherence weighted lower as supportive but secondary factors.

5. Results

We implemented an XGBoost-based rubric scoring system to evaluate the restaurant recommendation chatbots. The weighted scoring system (Factual: 50%, Relevance: 30%, Diversity/Coherence: 20% each) confirmed Embed_Bot's superiority, with final scores of **3.97 vs. 2.25** for KG_Bot.

Metric	KG Bot	Embed Bot	Winner
Relevance	1.970	3.985	Embed Bot
Factual	1.715	4.046	Embed Bot
Diversity	2.860	4.124	Embed Bot
Coherence	2.449	3.720	Embed Bot
OVERALL	2.248	3.969	Embed Bot

Fig.8. Bot Comparison Results

As shown in Figure 9, the embedding-based model outperformed the knowledge-graph model across all four criteria, with especially large gains in relevance and factual accuracy. This improvement can be attributed to embedding representations capturing richer semantic meaning, enabling closer alignment between user queries and restaurant reviews. While the KG-based approach relied heavily on structured entity matching—making it more prone to errors when queries contained noise or ambiguity—the embedding-based model was more flexible in interpreting intent and retrieving contextually suitable results. As a result, Embed_Bot not only provided factually correct recommendations but also delivered more varied and coherent outputs, demonstrating the advantage of semantic embeddings over purely symbolic graph structures for this task.

The model demonstrated reliable automated rubric evaluation, with hyperparameter tuning improving metrics across the board: MAE decreased from 0.369 to 0.353, RMSE improved from 0.699 to 0.650, and cross-validation MAE reduced from 0.548 to 0.543 (±0.394), indicating strong generalization without overfitting.

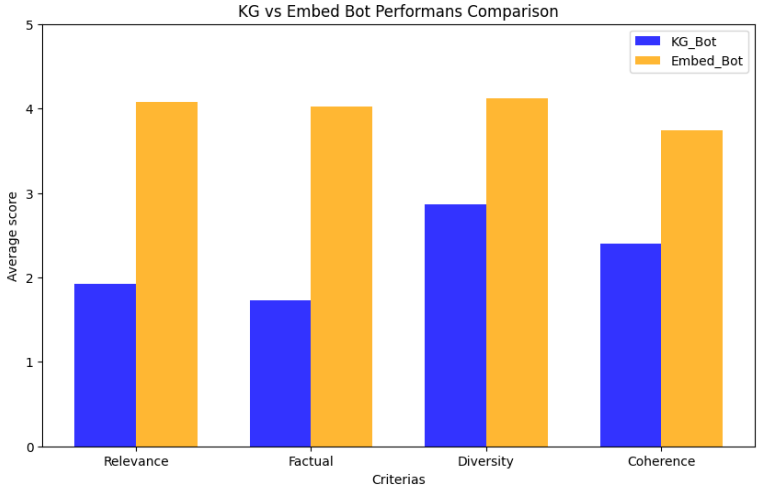

Fig.9. KG vs Embed BOT performance comparison chart

6. Conclusions

This study demonstrated how unstructured restaurant reviews can be transformed into actionable insights and personalized recommendations through an integrated NLP pipeline. By combining sentiment classification, aspect-based analysis, summarization, embeddings, and knowledge graph construction, we built a chatbot capable of delivering contextually relevant and quality-aware restaurant suggestions. The evaluation showed that the embedding-based approach consistently outperformed the knowledge-graph model, particularly in relevance and factual accuracy, underscoring the strength of semantic representations for noisy, user-generated text. Beyond restaurant recommendations, the proposed framework highlights the potential of LLM-powered pipelines for structuring large-scale feedback data and enabling more intelligent, user-centric decision support systems.

7. Future Work

Future work may focus on a few key areas. First, it may be beneficial to fine-tune ABSA models for Turkish to better capture local language nuances. Additionally, the system could be expanded to handle multimodal inputs, like images of dishes and menus. The chatbot may also be deployed as a scalable web service and tested against commercial recommender systems.
Another challenge is the limited availability of Turkish review data. While multilingual models like mBERT and XLM-RoBERTa help with cross-lingual adaptability, they still need fine-tuning with culturally relevant datasets. More efficient models may be developed to learn from sparse multilingual data while maintaining accuracy.

References

1. Cuizon, J., Simpao, J., & Ang, M. (2019). *Aspect-based sentiment analysis of restaurant reviews using hybrid classification methods*. Proceedings of the International Conference on Artificial Intelligence.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT, 4171–4186.
3. Hossain, E., Khan, M., & Rahman, M. (2020). *A survey on sentiment analysis in social media: Techniques, tools, and applications*. International Journal of Computer Applications, 975, 8887.
4. Hua, K., Xu, W., & Wang, H. (2024). *Aspect-based sentiment analysis: Advances, applications, and challenges*. Journal of Artificial Intelligence Research, 75, 145–178.
5. Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
6. Loesch, F., Kramer, M., & Schmidt, M. (2022). *Knowledge graphs in food recommender systems: Using embeddings for culinary substitutions*. Proceedings of the ACM Conference on Recommender Systems.
7. Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. Proceedings of EMNLP, 79–86.
8. Rahman, T., Ahmed, R., & Lee, J. (2025). *Instruction-tuned large language models for aspect-based sentiment analysis*. Proceedings of ACL, 2123–2137.
9. Wang, X., He, X., Cao, Y., Liu, M., & Chua, T. S. (2019). *KGAT: Knowledge graph attention network for recommendation*. Proceedings of KDD, 950–958.