

Early detection of food safety risks using BERT and large language models

Mohammed el Amin Gasbaoui, Soumia Benkrama, Mostefa Bendjima

Laboratory of TIT, Department of Mathematics and Computer Science, Faculty of Exact Sciences, Tahri Mohammed University, Bechar, Algeria

Article Info

Article history:

Received Oct 20, 2024

Revised Mar 20, 2025

Accepted Jul 2, 2025

Keywords:

Arabic sentiment analysis

Deep learning

Large language model

NLP

Parallel processing

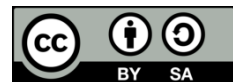
RESTful API

Software design

ABSTRACT

Sentiment analysis can be a powerful tool in safeguarding public health. This allows authorities to investigate and take action before a foodborne illness outbreak spreads. This paper introduces a novel system that proactively empowers restaurants to identify potential food safety hazards and hygiene regulation violations. The system leverages the power of natural language processing (NLP) to analyze Arabic restaurant reviews left by customers. By fine-tuning a pre-trained BERT mini-Arabic model on three targeted datasets: Sentiment Twitter Corpus, an Algerian Dialect dataset, and an Arabic restaurant dataset, the system achieves an impressive accuracy of 91%. Additionally, the system caters to spoken feedback by accepting audio reviews. We utilized Whisper AI for accurate text transcription, followed by classification using a fine-tuned Gemini model from Google on Algerian local comments and others generated using large language models (LLMs) through few-shot learning techniques, reaching an accuracy of 93%. Notably, both models operate independently and concurrently. Leveraging RESTful APIs, the system integrates the solved sub-solution from each microservice into a fusion layer for a comprehensive restaurant evaluation. This multifaceted approach delivers remarkable results for both modern standard arabic (MSA) and the Algerian dialect, demonstrating its effectiveness in addressing restaurant food safety concerns.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mohammed el Amin Gasbaoui

Laboratory of TIT, Department of Mathematics Computer Science, Faculty of Exact Sciences

Tahri Mohammed University

Independence Road B.P 417, Bechar 08000, Algeria

Email: gasbaoui.mohammedelamin@univ-bechar.dz

1. INTRODUCTION

Food safety is vital to public health, ensuring that food is free from harmful contaminants through proper handling, storage, and hygiene practices in production and preparation. International organizations like the Food and Drug Administration (FDA) and World Health Organization (WHO) have established strict regulations across the food supply chain to evaluate and control potential hazards [1].

Food safety monitoring involves conducting regular inspections to identify potential hazards, ensuring that procedures are correctly implemented and food safety regulations are properly followed [2].

The global demand for food increases the risk of foodborne diseases. Over 600 million cases occur annually, resulting in 420,000 deaths [3]. Food protection is essential in the restaurant industry as it directly affects public health. Proper practices in food handling, storage, and preparation are essential to prevent foodborne illnesses. Commitment to strict food safety regulations ensures that the meals served are safe for consumption, maintains customer trust, and protects the establishment from legal and reputational risks.

As artificial intelligence (AI) technologies advance, the creation and deployment of digital food systems are becoming more attainable. There is substantial interest in employing various AI applications, including machine learning models, natural language processing (NLP), and computer vision, to enhance food safety [4]. Machine learning has become an essential tool for tasks like data analysis, classification, and making predictions [5]. Analyzing the behavior of emotions, opinions, and feelings expressed in text about a specific entity or subject is known as opinion mining or sentiment analysis [6].

The rise of large language models (LLMs) has significantly advanced NLP, improving tasks like sentiment analysis [7], language generation, and machine translation. Through extensive pre-training on diverse datasets, LLMs capture nuanced language patterns and contextual meanings, leading to enhanced accuracy and efficiency in various applications.

In this paper, we aim to mitigate risks related to food and hygiene regulation violations in restaurants through sentiment analysis of customer reviews. We focused on both modern standard arabic (MSA) and the Algerian dialect by proposing a system composed of independent microservices that handle different user inputs, including text and audio. The outputs from each microservice are aggregated in the fusion layer to evaluate whether a restaurant is safe or not.

This paper's structure is as follows: section 2 reviews existing literature on sentiment analysis. Section 3 outlines our proposed system, which includes three main components. First, we describe the training process for our sentiment analysis model using three different datasets. Second, we detail the audio transcription process and the fine-tuning of Google's Gemini model. Third, we present the algorithms for the fusion layer, which integrates all solved subproblems for final scoring. Section 4 discusses our findings in detail. Finally, section 5 concludes by summarizing the insights gained from our research and suggesting potential directions for future work.

2. RELATED WORKS

Sentiment analysis in the AI field represents a transformative approach to understanding human emotions and opinions through text data. By leveraging NLP and machine learning techniques, sentiment analysis deciphers the underlying sentiments expressed in social media posts, reviews, and other textual content. In a recent study [8], a deep feed-forward neural network (DFFNN) was developed for recognizing emotions in Arabic speech, specifically categorizing them into three classes: happy, angry, and surprised. This study utilized the arabic natural audio dataset (ANAD) and examined a set of extracted features. The application of principal component analysis (PCA) for dimensionality reduction resulted in an accuracy of 98.56%. Additionally, the borderline-synthetic minority over-sampling technique (B-SMOTE) was employed to address the imbalanced dataset.

The authors in [9] developed a dialectal Arabic tweet act dataset by annotating a subset of the extensive arabic sentiment analysis dataset (ASAD) based on six speech act categories. They compared various Arabic BERT model variants and evaluated all models using a previously developed Arabic Tweet Act dataset (ArSAS). To address the class imbalance issue commonly observed in speech act problems, they implemented a transformer-based data augmentation model to generate a balanced proportion of speech act categories. The results indicate that the best-performing BERT model is the araBERTv2-Twitter model, which achieved a macro-averaged F1 score of 0.73 and an accuracy of 0.84. Additionally, the authors proposed an ensemble method, which is defined as a combination of multiple models to achieve better performance than any single model. This ensemble method outperformed the araBERTv2-Twitter model, with an accuracy of 0.85 and an F1 score of 0.74.

In the study [7], a comparison of various LLMs in handling nuanced and ambiguous scenarios was conducted. The researchers translated 20 scenarios across different cases into ten languages and predicted the associated sentiments. The findings indicate that ChatGPT 3.5, ChatGPT 4, and Gemini Pro generally provide accurate sentiment predictions in most ambiguous scenarios. However, these models frequently struggle to detect nuances such as irony or sarcasm. Additionally, they exhibit linguistic biases, associating sentiment with specific language families.

In [10], a sentiment analysis of Arabic restaurant customer reviews was conducted, focusing on four predefined aspects: price, cleanliness, food quality, and service. A dataset comprising 3,000 reviews was collected from the website Jeeran.com and labeled by 363 students from the University of Jordan. Four different models were employed for feature extraction: (1) term frequency (TF) only, (2) TF combined with Chi-Squared (Chi2), a filter-based feature selection approach that reduces the dimensionality of the feature space while preserving important information, (3) positive-negative-negation (PNN) lists only, and (4) TF-PNN-Chi2. The results indicated that the TF-PNN-Chi2 method yielded the best performance when Support Vector Machine (SVM) classifiers were applied.

The presented approach in [6] offers a hybrid method for sentiment analysis, integrating review-related and aspect-related features to create unique hybrid feature vectors (HFV) for each review. These vectors are then utilized for sentiment classification employing the deep learning classifier long short term memory (LSTM). The model was tested on three different datasets: SemEval-2014, Sentiment140, and STS-Gold. The results indicated that employing HFV+LSTM yielded average precision, average recall, and average F1-scores of 94.46%, 91.63%, and 92.81%, respectively, within the SemEval-2014 dataset.

Cross-modal BERT (CM-BERT) was proposed by [11], which utilizes both text and audio modalities to fine-tune the pre-trained BERT model. The authors developed a new masked multimodal attention mechanism that dynamically adjusts word weights by leveraging interactions between text and audio modalities. Two public sentiment analysis datasets, CMU-MOSI and CMU-MOSEI, were used to evaluate the multimodal approach. COVAREP, a feature extraction tool specifically designed for analyzing speech signals, was employed to extract audio features. COVAREP stands for “COntinuous VAlue Representation of Speech” and extracts various acoustic features from speech signals. The CM-BERT model created a new state-of-the-art result on the MOSI dataset, improving the performance on all the evaluation metrics.

Choosing the right sentiment analysis technique plays a crucial role in accurately gauging the sentiment of text data. This selection process often involves a trade-off between simplicity, interpretability, and the ability to capture complex nuances. Table 1 shows a comparison summarizing the strengths and weaknesses of several popular techniques.

Table 1. Comparison of the most NLP methods.

Technique	Description	Advantages	Disadvantages
- Bag of Words (BoW) [12]	- Represents text as a collection of word frequencies.	- Simple and easy to implement.	- Fails to capture word order and context.
- TF-IDF [13]	- Weighs words based on their importance in a document relative to the corpus.	- Provides better differentiation of words' significance.	- Lacks context understanding.
- Word2Vec [14]	- Word embedding technique that represents words as vectors in a high-dimensional space. Words with similar meanings have similar vector representations.	- Enhances context understanding.	- Requires large datasets for training. Performance can be affected by data quality.
- GloVe (Global Vectors for Word Representation) [15]	- Uses global word co-occurrence statistics to produce embeddings.	- Captures semantic relationships between words, often outperforming Word2Vec in specific tasks.	- Requires large datasets for training. May not capture all nuances of word meaning.
- BERT (Bidirectional Encoder Representations from Transformers) [16]	- Utilizes bidirectional transformer architecture for deep contextual understanding.	- Provides advanced context understanding.	- Computationally intensive and requires substantial memory.
- ALBERT (A Lite BERT) [17]	- Shares parameters across layers to reduce model size and improve efficiency.	- Reduces model size and improves efficiency while maintaining performance.	- May have slightly lower accuracy compared to BERT for complex tasks.

3. PROPOSED APPROACH

The proposed architecture in Figure 1 is divided into microservices, each handling specific tasks independently. One microservice processes text comments and emojis. Using a pre-trained BERT mini Arabic model, the text input is classified into three categories (positive, neutral, and negative). The classification output is then compared with the user's input emoji (happy, neutral, or unhappy). The user must input both text and one emoji. The evaluation phase assigns a score to the user's review, a score is awarded if the emoji aligns with the text classification. Reviews with inconsistent text and emoji inputs are discarded. Another microservice processes optional audio input from the user. The audio is transcribed using OpenAI's Whisper model, and the resulting text is classified using the pre-trained Gemini model. These microservices are loosely coupled, allowing them to execute independently and in parallel. The results from each microservice are stored in a remote database via a REST API.

3.1. Text model classification for arabic sentiment analysis

We combined three datasets: arabic restaurant reviews from hugging face [18], the Arabic Sentiment Twitter Corpus from Kaggle [19], and the Algerian Dialect dataset [20]. The Algerian Dialect dataset and the Arabic Sentiment Twitter Corpus contain comments in various fields, we employed keyword extraction to identify reviews related to food, restaurants, and hygiene. We collected over 100 food and restaurant-related keywords, such as (‘شهي’ - Delicious ‘وجبة’ - Meal ‘مطبخ’ - Kitchen ‘مأكولات’ - Cuisine ‘متسخ’ - Dirty...). Our collection comprised more than 54,000 text comments labeled as either positive (1) or negative (0). During the preprocessing step, while the majority of Algerians use Arabic letters, some prefer Latin characters. Therefore, we removed all Latin characters, Arabic punctuation like Shadda (ّ), Fatha (َ), Damma (ُ), and special characters like (? , #, \$,...). Additionally, we eliminated elongations (a form of Arabic diacritics) such as (“ء”, “ى”, “ة”, “و”, “و”) and stopwords using the NLTK library for the Arabic language. Machine learning models trained on imbalanced datasets may develop biases toward the majority class. For that, we balanced the data to prevent bias. We obtained a total of 54,164 text samples, evenly distributed between the two classes, with 27,082 positive and 27,082 negative samples.

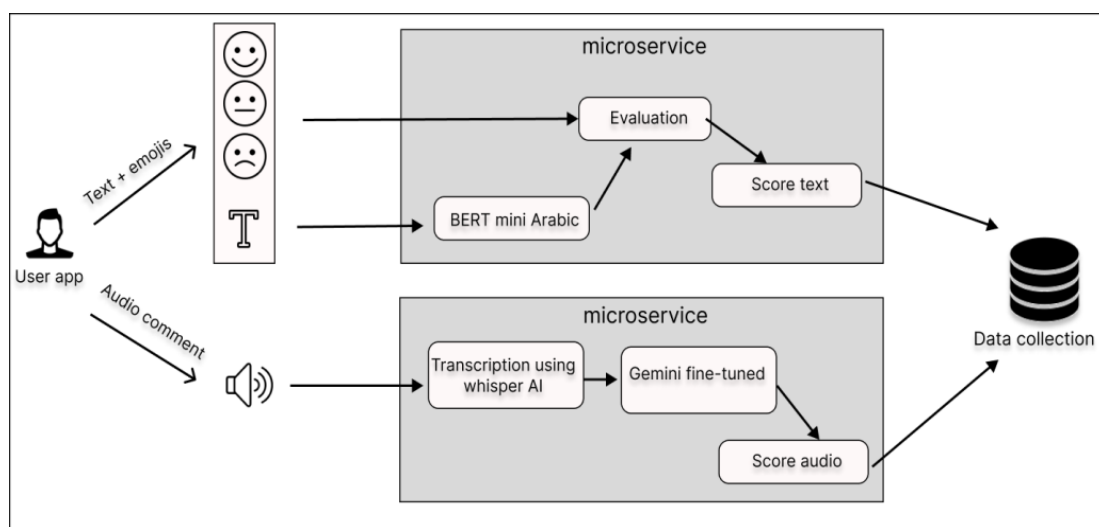


Figure 1. System architecture for arabic sentiment analysis in customer restaurant reviews

To fine-tune a BERT model for Arabic sentiment analysis, we used the ‘asafaya/bert-mini-arabic’ model [21], along with its corresponding tokenizer provided by the Hugging Face Transformers library. Additionally, specific tokens such as (CLS) to indicate the start of an input sequence and (SEP) for separation at the end are added to each sentence. All sentences are then standardized to a uniform length through padding and truncation. 80% (43,331 samples) are allocated for training, 10% (5,416 samples) for validation, and another 10% (5,417 samples) for testing purposes. The pre-trained model requires two crucial inputs: ‘input ids,’ which represent numeric vectors of each word in the input sequence, and ‘attention masks’ indicating which tokens the model should prioritize or ignore. We set the maximum length for each sequence input to 64 characters, padding shorter tokenized reviews with zeros and truncating longer ones.

In BERT, input text is tokenized and then transformed into dense vectors through a series of neural network layers. These dense vectors are the hidden states that capture the input text’s semantic meaning. The hidden size of our classifier is set to 256, and the number of output labels is 2, corresponding to positive and negative sentiments. The inclusion and range of a third class (neutral) will be addressed in the following section. We employed the AdamW optimizer, a variant of the Adam optimizer known for its Weight Decay regularization technique. Weight decay penalizes large weights in the model, effectively preventing overfitting and enhancing generalization performance.

3.2. Audio speech transcription

Audio speech transcription is the process of converting spoken language in an audio recording into written text. For that, we employed the famous pre-trained model Whisper AI created by OpenAI. Whisper is a highly capable speech recognition model that is trained on a vast collection of diverse audio clips. we used

the medium model to interact with Arabic languages. Using the whisper AI model, we directly transcribe the speech into text.

3.3. Gemini model tuning for arabic sentiment classification

We gathered approximately 207 reviews covering various scenarios related to food safety and hygiene. Each review was meticulously labeled into three categories: positive, neutral, and negative. For instance, the comment “We watched the match on a giant TV, and it was great while we were waiting for the order from the waiter restaurant” would be considered positive in the context of entertainment. However, in our study, it is labeled as neutral because our primary goal is to identify any indications of foodborne illness and assess the state of hygiene.

In addition to collecting data from the local region, including MSA and Algerian dialects, we enriched the dataset by utilizing LLMs like Gemini, Claude and ChatGPT. Using few-shot learning techniques, we provided specific prompts to direct the generated output and ensured the LLMs used the Algerian Arabic dialect. A few-shot techniques involve supplying the model with a small number of examples or prompts to guide its understanding and response generation for a specific task. Table 2 shows some sample dataset.

Table 2. Sample dataset using collected local data and LLMs.

Reviews	Sentiment analysis
المأكلة كانت بنيفة بزاف! وحتى الخدمة كانت رائعة، ماشاء الله. حتماً نعاود نجيو مرة أخرى. The food was really delicious! And the service was great, God willing. We'll definitely come back again.	Positive
يا إلهي ، هذا المطعم خرافي! الطاجين اللذيذ كنا ناكلوه كان لذيذ بجنون، والخدمة سريعة وودودة. نصبحك تجربوه، ما راحش تندموا! Oh my God, this restaurant is amazing! The tagine we ate was incredibly delicious, and the service was fast and friendly. I highly recommend you try it, you won't regret it!	Positive
الموقع للمطعم جيد، قريب من الجامعة. لكن المكان صغير شوي، صعب نجد طاولة خالية. The restaurant's location is good, close to the university. But the place is a bit small, it's hard to find an empty table.	Neutral
المطعم جديد، الديكور حديث وجميل. لكن قائمة الطعام محدودة، ما فيهاش خيارات كثيرة. The restaurant is new, the decor is modern and beautiful. But the menu is limited, there aren't many options.	Neutral
المكان غير نظيف، كان هناك فوضى وأوساخ على الطاولات. The place is not clean, there was a mess and dirt on the tables.	Negative
لم أكن مرتاحاً لأن المطعم كان غير نظيف والطعام لم يعجبني. I wasn't comfortable because the restaurant was not clean, and I didn't like the food.	Negative

We fine-tuned the existing model “gemini-1.0-pro-001” from Google. This process involves adapting the pre-trained large language model to a specific task by providing it with a domain-specific dataset and adjusting parameters to enhance its performance on the new task. Gemini models can be customized to excel at specialized tasks through fine-tuning, training them on your data not only enhances their performance but also streamlines interactions by reducing the needed context and speeding up response times [22]. We divided the dataset into 80% for training and 20% for testing. We set the temperature to 0.9, as this parameter controls the randomness of the model's predictions. We chose top_p=1.0, which determines the cumulative probability threshold for token selection. This ensures the model considers only the top few tokens, leading to a more focused and predictable output. We configured the model with 7 epochs, a batch size of 4, and a learning rate of 0.001 for the training phase.

3.4. Fusion layer implementation for the overall restaurant assessment

The overall system is divided into microservices. Users must send a text input with at least one emoji (happy, neutral, or unhappy) and have the option to share an opinion about the eatery as an audio input. Communication is based on a RESTful API (see Figure 2). The first microservice has the following endpoint: (id-request, id-restaurant, emoji, text-input). The second microservice has: (id-request, id-restaurant, audio-input). The id-request is used to track comments, meaning that the same id-request received by both the first and second microservices is generated from the same user. The third microservice gathers the output of each solved sub-problem with an endpoint (id-request, id-restaurant, score text or audio). This microservice also tracks requests from restaurant customers for the next review.

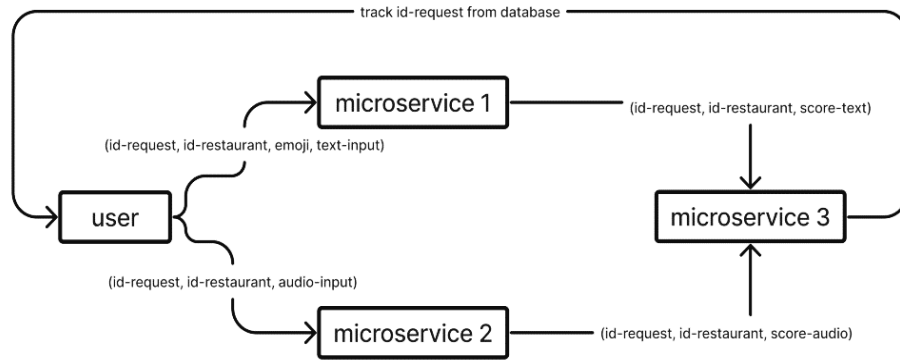


Figure 2. RESTful API communication between the microservices.

The algorithm in Figure 3 shows the evaluation function for both text and audio input. The evaluateText function, Figure 3(a), assigns a score to each comment based on its sentiment. For text input, the associated emoji must match the predicted sentiment. If there is a discrepancy between the emoji and the predicted sentiment, a score of -1 is given, and the comment is ignored. The same logic applies to the evaluateAudio function in Figure 3(b), which processes audio as input.

```

Function evaluateText (id-request, id-restaurant, emoji, text-input)
begin //microservice 1 algorithm
  if (emoji= "happy" and predicted-text= "positive") return 0.5
  else if (emoji= "unhappy" and predicted-text= "negative") return -0.5
  else if (emoji= "neutral" and predicted-text= "neutral") return 0
  else return -1
end
  
```

(a)

```

function evaluateAudio (id-request, id-restaurant, audio-input)
begin //microservice 2 algorithm
  text=transcription(audio-input)
  if (text= "positive") return 0.5
  else if (text= "negative") return -0.5
  else if (text= "neutral") return 0
end
  
```

(b)

Figure 3. The algorithms for (a) text evaluation algorithm (b) audio evaluation algorithm

We used a dictionary data structure based on key-value pairs. DT (dictionary text) is a dictionary variable where the key is the request id and the value is a tuple containing the restaurant id and the text score (id-restaurant, score-text). Similarly, DA (dictionary audio) is a dictionary like DT, but its value contains the audio score (id-restaurant, score-audio) instead of the text score. In (1) shows the global score of a restaurant identified by its id. The overall score of a restaurant is the sum of the text and audio scores, computed by iterating over all available requests and summing the scores where the current restaurant id matches the restaurant id being evaluated. For a rigorous evaluation, we sum the text and audio scores only if both have the same sentiment. In the case of a contradiction, we avoid summation. st and sa in the equation represent the score text and score audio respectively.

$$\text{Score}_{\text{id-rest}} = \sum_{\substack{\text{id-req}=0 \\ \text{DT}_{\text{id-req}}[\text{id-rest}]=\text{id-res} \\ \text{and} \\ \text{DA}_{\text{id-req}}[\text{id-rest}]=\text{id-res} \\ \text{and} \\ \text{DT}_{\text{id-req}}[\text{st}]=\text{DA}_{\text{id-req}}[\text{sa}]}}^{\text{max-id-req}} \text{DT}_{\text{id-req}}[\text{st}] + \text{DA}_{\text{id-req}}[\text{sa}] \quad (1)$$

4. RESULTS AND DISCUSSION

4.1. Text model classification results

We employed a single Tesla P100-PCIE-16GB GPU for our experiments. The hyperparameters used were a learning rate (lr) of 5e-5, a batch size of 16, and training for 4 epochs. Additionally, we set the epsilon (eps) value to 1e-8, which is a small constant added to the denominator of the update step formula to prevent division by zero. The results of our experiments are presented in Table 3. The validation accuracy reached 90% with a validation loss of 0.26 during the fourth epoch.

Table 3. Performance of the Arabic model on training and validation dataset

epochs	Train loss	Val loss	Val accuracy
1	0.378216	0.316435	87.05
2	0.257231	0.307856	89.13
3	0.203590	0.286417	89.86
4	0.169917	0.257172	90.01

The model achieved 91% accuracy on the testing set, with a precision, F1-score, and recall of 90%. When tested on 100 local reviews, the fine-tuned model achieved its best performance by setting the probability range for the neutral class to]0.2, 0.8[, the positive class to [0.8, 1], and the negative class to [0, 0.2]. The model correctly predicted 91% of positive and negative reviews and 88% of neutral ones. BERT-mini-Arabic was selected for its ability to capture complex language patterns and contextual meanings in Arabic while its lightweight architecture ensures fast response times.

4.2. Gemini tuning classification results

Figure 4 shows the loss function of a fine-tuned Gemini model during training after 7 epochs and a batch size of 4. The model is rapidly learning from the training data, which is typical for LLMs that can quickly absorb patterns and relationships in vast datasets. By epoch 3, the loss curve plateaus and oscillates around a low value. This indicates that the model has reached a point where it is no longer making significant progress with each additional epoch. The accuracy reached 93% on the testing subset.

4.3. Result of fusion layer implementation

Specifying the language as “ar” instead of allowing Whisper AI to detect it automatically leads to faster transcription. This is because the model can bypass the language detection step and directly focus on transcribing the audio. This reduction in processing steps results in quicker response times, enabling a more efficient transcription process.

The proposed algorithm operates with a time and space complexity of $O(RQ)$, where R represents the size of the restaurant list containing properties such as restaurant id, name, and global score, and Q denotes the size of the requests list. The global score reflects the total evaluation of each restaurant after looping over all available requests. Thus, the algorithm’s limitation can be succinctly expressed as $RQ < 10^8$.

We gathered approximately 60 entries, some including text comments and audio input and some without audio. We compared the actual scores with the predicted scores for five restaurants. Figure 5 illustrates the comparison results. For restaurants one and four, the actual scores matched the predicted scores, indicating a positive status with no signs of foodborne illness or hygiene regulation breaches. In the case of restaurant two, there was a slight discrepancy between the actual and predicted scores due to the misclassification of two comments. However, the final assessment for this restaurant remains positive, posing no threat to food safety. The negative scores for restaurants three and five indicated potential risks and hygiene regulation breaches, necessitating rapid intervention. To ensure accurate decisions for each eatery, evaluations are considered only after the restaurant has received more than three reviews. The proposed system delivered promising results, emphasizing its accuracy and efficiency through task parallelization. Our solution demonstrated its potential to address food safety concerns comprehensively.

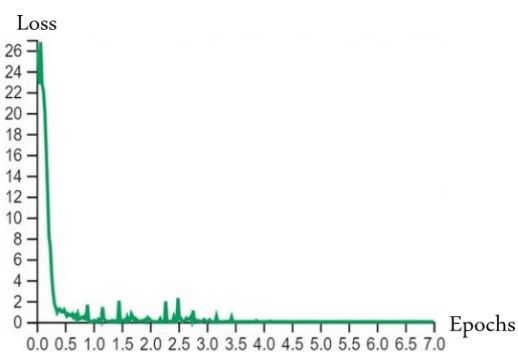


Figure 4. The loss function of the fine tuned gemini model across the epochs.

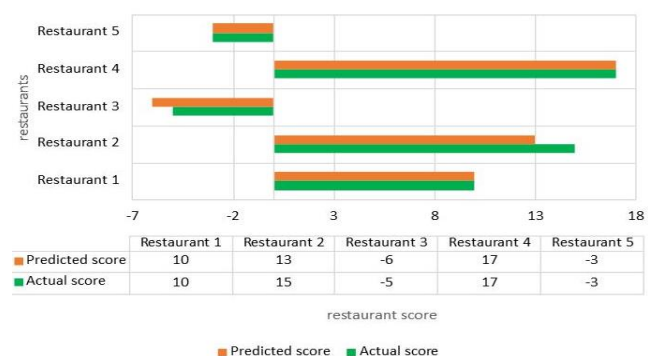


Figure 5. The comparison results of actual scores versus predicted scores of five restaurants.

4.4. The system's limitations and the comparison results with previous studies

Most existing research overlooks the food safety domain, underscoring the need for a comprehensive system to detect risks that threaten consumer safety. The proposed system, compared in Table 4, employs a parallel processing architecture, allowing each microservice to scale independently and ensuring fault isolation, which improves both scalability and response times without compromising overall performance. The system faces challenges in accurately classifying dialect words, especially when reviews use Latin characters, and struggles with variations in accents. Whisper AI's limitations with the Algerian dialect lead to occasional transcription inaccuracies. Furthermore, the use of RESTful API communication can increase latency and network overhead.

Table 4. Comparison results with the previous studies

Reference	Dataset	Size	Design system	Used method	Accuracy
[20]	Algerian dialect dataset	45000	monolithic architecture	BERT	81.74%
[10]	Collected data (jeeran website)	3000	monolithic architecture	Machine learning (SVM) + domain specific dictionaries and sentiment word lists.	84.47%
[23]	ATSAD	36000	monolithic architecture	distant supervision and self-training approaches for labels. machine learning word and character grams features.	86%
[24]	collected data (different newspaper websites).	1000	monolithic architecture	SVM + Random Sub Space (RSS) and genetic algorithm.	85.99%
[25]	collected data (various areas of life).	7698	monolithic architecture	lexicon-based approach	79.13%
[26]	TWIFIL (collected tweets)	9000	monolithic architecture	CNN	76%
[27]	collected data (cafes and restaurants from qassim region).	1785	monolithic architecture	Machine learning (SVM)+ 10-fold cross-validation	92%
[28]	AraMA (Google Maps reviews for restaurants in Riyadh).	10750	monolithic architecture	Machine learning (SVC kernel linear model)	66.67% in term of aspect category.
				Machine learning (SGD)	56.49 % for positive sentiment
				Machine learning (SVC kernel linear model)	75.70% for negative sentiment
				Machine learning (SVC kernel linear model)	79.33% For Conflict and Neutral Sentiment
the proposed system	combined datasets: - Arabic restaurant reviews - Sentiment Twitter Corpus - Algerian Dialect dataset	54164	Microservice-based architecture utilizes parallel processing with RESTful API communication.	bert-mini-arabic	91%
	collected data (local data+ few shot learning using LLM)	207		fine-tuned Gemini model from Google	93%

5. CONCLUSION

The proposed system demonstrates high accuracy and efficiency in addressing food safety concerns by analyzing restaurant reviews in both MSA and the Algerian dialect. By fine-tuning the BERT mini Arabic model and Gemini model, we achieved accuracies of 91% and 93%, respectively. The system consists of loosely coupled, parallel-operating microservices that are integrated through a RESTful API to provide comprehensive restaurant evaluations. Future work will focus on improving accuracy and expanding the system's applicability to other dialects and regions.

ACKNOWLEDGEMENTS

We reiterate that this article has not been published anywhere and is not sponsored by any particular organization. No funding was received for conducting this study.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohammed el Amin Gasbaoui	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Soumia Benkrama		✓		✓		✓				✓	✓	✓		
Mostefa Bendjima		✓		✓							✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

REFERENCES

- [1] G. Makridis, P. Mavrepis, and D. Kyriazis, "A deep learning approach using natural language processing and time-series forecasting towards enhanced food safety," *Machine Learning*, vol. 112, no. 4, pp. 1287–1313, Apr. 2023, doi: 10.1007/s10994-022-06151-6.
- [2] X. Wang, Y. Bouzembrak, A. G. J. M. O. Lansink, and H. J. van der Fels-Klerx, "Application of machine learning to the monitoring and prediction of food safety: A review," *Comprehensive Reviews in Food Science and Food Safety*, vol. 21, no. 1, pp. 416–434, Jan. 2022, doi: 10.1111/1541-4337.12868.
- [3] Q. Zhou, H. Zhang, and S. Wang, "Artificial intelligence, big data, and blockchain in food safety," *International Journal of Food Engineering*, vol. 18, no. 1, pp. 1–14, Jan. 2022, doi: 10.1515/ijfe-2021-0299.
- [4] C. Qian, S. I. Murphy, R. H. Orsi, and M. Wiedmann, "How can AI help improve food safety?," *Annual Review of Food Science and Technology*, vol. 14, no. 1, pp. 517–538, Mar. 2023, doi: 10.1146/annurev-food-060721-013815.
- [5] G. Mohammed El Amin, B. Soumia, and B. Mostefa, "Water quality drinking classification using machine learning," in *2024 2nd International Conference on Electrical Engineering and Automatic Control, ICEEAC 2024*, May 2024, pp. 1–6, doi: 10.1109/ICEEAC61226.2024.10576539.
- [6] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," *Journal of Big Data*, vol. 10, no. 1, p. 5, Jan. 2023, doi: 10.1186/s40537-022-00680-6.
- [7] A. Buscemi and D. Proverbio, "ChatGPT vs Gemini vs LLaMA on multilingual sentiment analysis," *ArXiv*, 2024, [Online]. Available: <http://arxiv.org/abs/2402.01715>.
- [8] E. Abdelmaksoud, "Arabic automatic speech recognition based on emotion detection," *The Egyptian Journal of Language Engineering*, vol. 8, no. 1, pp. 17–26, Apr. 2021, doi: 10.21608/ejle.2020.49690.1016.
- [9] K. Alshehri, A. Alhothali, and N. Alowidi, "Arabic tweet act: a weighted ensemble pre-trained transformer model for classifying arabic speech acts on Twitter," *ArXiv preprint*, 2024, [Online]. Available: <http://arxiv.org/abs/2401.17373>.
- [10] F. Al-Smadi, B. Al-Shboul, D. Al-Darras, and D. Al-Qudah, "Aspect-based sentiment analysis of arabic restaurants customers' reviews using a hybrid approach," in *ACM International Conference Proceeding Series*, Oct. 2022, pp. 123–128, doi: 10.1145/3508397.3564834.
- [11] K. Yang, H. Xu, and K. Gao, "CM-BERT: cross-modal BERT for text-audio sentiment analysis," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 521–528, doi: 10.1145/3394171.3413690.
- [12] S. Schrauwen, "Machine learning approaches to sentiment analysis using the dutch netlog corpus," 2010.
- [13] S. Singh, K. Kumar, and B. Kumar, "Sentiment analysis of twitter data using TF-IDF and machine learning techniques," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, May 2022, pp. 252–255, doi: 10.1109/COM-IT-CON54601.2022.9850477.
- [14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

- [15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, pp. 1532–1543, doi: 10.3115/v1/d14-1162.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: a lite BERT for self-supervised learning of language representations," *ArXiv*, 2020, doi: 10.48550/arXiv.1909.11942.
- [18] H. ElSahar and S. R. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9042, 2015, pp. 23–34.
- [19] M. Saad, "Arabic sentiment twitter corpus," 2019. <https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus> (accessed Feb. 09, 2024).
- [20] Z. Benmounah, A. Boulesnane, A. Fadheli, and M. Khial, "Sentiment analysis on algerian dialect with transformers," *Applied Sciences*, vol. 13, no. 20, p. 11157, Oct. 2023, doi: 10.3390/app132011157.
- [21] A. Safaya, M. Abdullatif, and D. Yuret, 'KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media', in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 2054–2059. doi: 10.18653/v1/2020.semeval-1.271.
- [22] "Tune Gemini Pro in Google AI Studio or with the Gemini API," 2024. <https://developers.googleblog.com/en/tune-gemini-pro-in-google-ai-studio-or-with-the-gemini-api/> (accessed Jul. 06, 2024).
- [23] K. A. Kwaik, S. Chatzikyriakidis, S. Dobnik, M. Saad, and R. Johansson, 'An Arabic Tweets Sentiment Analysis Dataset (ATSAD) using Distant Supervision and Self Training'.
- [24] A. Ziani, N. Azizi, D. Zenakhra, S. Cheriguene, and M. Aldwairi, 'Combining RSS-SVM with genetic algorithm for Arabic opinions analysis', *IJISTA*, vol. 18, no. 1/2, p. 152, 2019, doi: 10.1504/IJISTA.2019.097754.
- [25] M. Mataoui, O. Zelmati, and M. Boumechache, 'A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic', *RCS*, vol. 110, no. 1, pp. 55–70, Dec. 2016, doi: 10.13053/rcs-110-1-5.
- [26] L. Moudjari, K. Akli-Astouati, and F. Benamara, "An algerian corpus and an annotation platform for opinion and emotion analysis," *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 1202–1210, 2020.
- [27] L. M. Alharbi and A. M. Qamar, "Arabic sentiment analysis of eateries' reviews: qassim region case study," in *Proceedings - 2021 IEEE 4th National Computing Colleges Conference, NCCC 2021*, Mar. 2021, pp. 1–6, doi: 10.1109/NCCC49330.2021.9428788.
- [28] A. AlMasaud and H. H. Al-Baity, "AraMAMS: Arabic multi-aspect, multi-sentiment restaurants reviews corpus for aspect-based sentiment analysis," *Sustainability (Switzerland)*, vol. 15, no. 16, p. 12268, Aug. 2023, doi: 10.3390/su151612268.

BIOGRAPHIES OF AUTHORS



Mohammed El Amin Gasbaoui    is a Ph.D. researcher who earned his master's degree from the University of Tahri Mohammed Bechar, Algeria. His research interests include machine learning, sentiment analysis, food safety, and big data. He can be contacted at email: gasbaoui.mohammedelamin@univ-bechar.dz.



Dr. Soumia Benkrama    is Lecturer Class 'A' professor in the Computer Science Department at the Faculty of Exact Sciences, University Tahri Mohamed-Bechar, Algeria. She earned her master's and Ph.D. in computer science from the University of Science and Technology Mohamed Boudiaf, Oran, Algeria. Her research focuses on image processing, machine learning, pattern recognition, and computational intelligence. She can be contacted at email: benkrama.soumia@univ-bechar.dz.



Dr. Mostefa Bendjima    is Lecturer Class 'A' professor in the computer science department at the faculty of exact sciences, University Tahri Mohamed-Bechar, Algeria. His research areas are wireless sensor network, agent system, intelligent communication, attack detection, and artificial intelligence. He can be contacted at email: bendjima.mostefa@univ-bechar.dz.