Prompt Engineering for Sentiment Analysis in Tourism: The Case of Mexican Pueblos Mágicos

Federico Sandoval[1,*]

[1]Algiedi Solutions, Cholula de Rivadavia, Puebla, México.

Abstract

The rise of large language models (LLMs) has enabled new paradigms for performing natural language processing tasks without the need for fine-tuning. Among these paradigms, prompt engineering has emerged as a key technique for adapting generic models to specific domains. In this study, we explore the potential of prompt-based methods for sentiment analysis in the tourism sector, focusing on Spanish-language reviews of destinations within the "Pueblos Mágicos" program in Mexico. We design and evaluate a set of carefully crafted prompts using both zero-shot and few-shot settings, targeting various commercial LLMs. Our results show that effective prompt design can yield competitive performance for polarity classification without requiring extensive training, and that specific linguistic cues related to hospitality and culture significantly affect model behavior. This work offers insights into the viability of prompt engineering for resource-constrained applications in domain-specific sentiment analysis, particularly in underrepresented languages like Spanish.

Keywords

Prompt Engineering, Sentiment Analysis, Mexican Tourism, Pueblos Mágicos, NLP, Large Language Models

1. Introduction

The exponential growth of user-generated content across digital platforms has radically transformed how travelers share, assess, and access tourism experiences. Online reviews—ranging from brief social media comments to detailed feedback on platforms like TripAdvisor or Google Reviews—have become indispensable for understanding customer satisfaction, service quality, and destination reputation [1, 2, 3]. These textual narratives offer rich, fine-grained insight into the tourist journey, encompassing emotional reactions, perceived authenticity, and even socio-political impressions of a place [4, 5]. In Mexico, where tourism is one of the primary engines of economic development, harnessing such data is critical for both public policy and private-sector competitiveness [6, 7].

Sentiment analysis, as a core task within natural language processing (NLP), has become a fundamental tool in mining this content. It allows institutions and businesses to gauge the affective tone of public opinion, thereby guiding service improvement, promotional strategies, and destination branding. However, deploying sentiment analysis in practice—especially for Spanish-language content—poses persistent challenges, such as limited resources, informal or noisy language, and the lack of annotated data across domain-specific subregions [8].

The Rest-Mex shared task, launched in 2021, has emerged as a benchmark for Spanish-language sentiment analysis in tourism. Initially designed to evaluate models for satisfaction prediction and polarity classification from tourist reviews [9], it has since evolved to encompass more complex tasks. The 2022 edition, for example, included epidemic status classification from news articles [10, 11], while in 2023 the challenge expanded geographically to include Cuban and Colombian data, introducing clustering-based tasks for opinion mining [12]. Yet, across these editions, the backbone of the evaluation has remained the ability to analyze sentiment and classify service types in multilingual and multicultural tourist data.

In the 2025 version of Rest-Mex, a new dimension was added by including geospatial granularity via the identification of "Pueblos Mágicos"—a designation by the Mexican government for towns with

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

fsandoval@algiedi.com.mx (F. Sandoval)

historical, cultural, or natural significance. Participants must now determine not only the sentiment and service category of a review but also its geographic anchor among 40 possible towns [13, 14]. This task amplifies the complexity of review classification, as it demands awareness of subtle contextual cues that might imply a specific location, without being explicitly stated.

Traditionally, addressing such challenges in NLP has involved pretraining and fine-tuning large language models (LLMs) like BERT, RoBERTa, or BETO—the latter being a Spanish-specific adaptation of BERT [11, 15]. Fine-tuning provides strong task performance, especially in resource-rich scenarios where labeled data is abundant. However, this approach can be computationally intensive and inflexible when rapid domain adaptation is required.

Recently, a paradigm shift has emerged with the advent of *prompt engineering*, wherein LLMs are guided using structured textual prompts rather than being retrained for each new task. This zero-shot or few-shot prompting strategy leverages the knowledge already encoded in massive foundation models, enabling them to generalize across tasks and domains with minimal adaptation. This shift is especially promising for low-resource settings and underrepresented languages, such as Spanish in Latin American contexts.

Prompt engineering reframes traditional NLP challenges as controlled language generation problems. Instead of optimizing parameters through gradient descent, the researcher optimizes the phrasing, structure, and examples included in the prompt. When carefully crafted, prompts can elicit accurate sentiment predictions, detect geographic indicators, and interpret informal or figurative expressions—all without modifying the internal weights of the model. This makes prompt engineering particularly suitable for domains like tourism, where language is diverse, creative, and rapidly evolving.

In this paper, we explore how prompt engineering can be leveraged to perform sentiment analysis on Spanish-language tourist reviews, with a specific focus on the "Pueblos Mágicos" of Mexico. We construct and evaluate a set of prompts across various LLMs using zero-shot and few-shot techniques. Our goal is to assess whether prompt-based sentiment classification can serve as a competitive alternative to fine-tuned models in domain-specific, multilingual scenarios. By examining model outputs across different prompt formulations, we aim to identify linguistic patterns that enhance performance and gain insights into how large-scale models interpret cultural and contextual nuances within Mexican tourism discourse.

This work contributes to a growing body of research that seeks to democratize access to NLP tools in low-resource languages and specialized domains. It also bridges methodological innovation with practical application, offering a framework for public agencies, tourism observatories, and small businesses to derive actionable sentiment insights without requiring high-end infrastructure or deep technical expertise.

2. State of the Art

The task of sentiment analysis in tourism intersects multiple domains of natural language processing, particularly those concerned with affective computing, opinion mining, and socio-geographic text understanding. In recent years, a substantial body of literature has explored how computational techniques can extract emotional and evaluative signals from tourist narratives, typically through supervised learning approaches on annotated corpora. Most early methods relied on bag-of-words or lexical approaches, but the field has shifted toward deep learning models, especially Transformer-based architectures, for their superior ability to model context, nuance, and sequential dependencies.

In the tourism domain specifically, sentiment analysis has proven to be a valuable tool for capturing visitor satisfaction, identifying pain points in services, and even detecting seasonality in perception trends [1, 5]. However, the subjective nature of tourist reviews—often shaped by personal expectations, cultural background, or transient experiences—makes them challenging to process reliably. These texts are often informal, metaphorical, or hyperbolic, which further complicates rule-based or lexicon-driven approaches.

To address this, the community has embraced the use of pre-trained models fine-tuned for specific

sentiment classification tasks. One such model is BETO [**?**], a Spanish-language version of BERT, which has demonstrated competitive results in various tasks including sentiment analysis, named entity recognition, and text classification [16]. Studies such as [3, 17] have successfully applied fine-tuned BETO-based classifiers in tourism datasets, reporting substantial performance gains over traditional baselines.

The Rest-Mex shared task series has further fueled the development of Spanish NLP for tourism by offering large-scale, annotated corpora of reviews and structured challenges with multiple classification tracks. Each year since 2021, Rest-Mex has introduced new complexities: from predicting satisfaction scores and sentiment polarity, to classifying COVID-19 status from news [9, 10], and even performing thematic clustering across multiple countries [12]. While most top-performing systems rely on fine-tuned Transformers, they also require significant infrastructure and domain adaptation, which may be impractical in real-world deployment scenarios, particularly for small enterprises or local governments.

This has given rise to growing interest in *prompt engineering* as a lightweight, scalable alternative to fine-tuning. Prompt engineering refers to the art of crafting input queries that guide large language models (LLMs) toward specific outputs, leveraging their pre-trained knowledge without altering their weights. Models like GPT-3, GPT-4, PaLM, and LLaMA have demonstrated that prompt-based learning can be effective across a wide range of tasks, including sentiment analysis, summarization, and question answering [18, 19].

In the context of sentiment analysis, prompt engineering enables users to frame polarity detection as a natural language task. For example, instead of training a model to classify the sentence "The hotel was amazing" as positive, a prompt-based system might be asked: “What is the sentiment of the following review?” followed by the review text. The model, pre-trained on diverse data, can then generate a response such as “positive” without the need for labeled training data.

This zero-shot or few-shot paradigm is particularly attractive in Spanish and other under-resourced languages, where annotated corpora are limited and task-specific models are scarce. Moreover, prompt engineering allows rapid prototyping and domain testing by adjusting linguistic templates rather than retraining full models. Techniques such as chain-of-thought prompting, few-shot examples, and instruction tuning have further expanded the flexibility of this approach.

Despite its promise, prompt engineering remains an emerging field with open questions about optimal prompt design, robustness across domains, and cultural or linguistic biases embedded in LLMs. Some studies have shown that prompt phrasing significantly affects outcomes, especially in subjective tasks like sentiment analysis. Additionally, most existing benchmarks and evaluations remain concentrated in English, creating an urgent need to assess prompt engineering in Spanish-language domains.

In this work, we aim to bridge this gap by applying prompt engineering to a real-world Spanish-language sentiment analysis task in tourism. Our focus on the "Pueblos Mágicos" dataset from Rest-Mex 2025 provides a rigorous and culturally grounded testbed for evaluating prompt design strategies in a multilingual, domain-specific setting.

3. Methodology

Our methodological framework is grounded in the evaluation of prompt-based sentiment analysis using large language models (LLMs), specifically within the context of Spanish-language tourism reviews. Unlike prior approaches that involve fine-tuning model parameters, our study investigates how carefully crafted prompts can direct the behavior of general-purpose LLMs to perform polarity classification without explicit model training. We structure our methodology into three main stages: dataset preparation, prompt design, and model evaluation.

3.1. Dataset Description

We conducted our experiments using the official training set provided by the *Rest-Mex 2025* shared task [13, 14], a large-scale benchmark dataset designed to advance sentiment and thematic analysis in the

tourism domain. The dataset consists of over 208,000 user-submitted reviews in Spanish, gathered from prominent tourism platforms such as TripAdvisor, Booking.com, and Google Reviews.

Each review in the dataset is annotated with three categorical labels: sentiment polarity (ranging from 1 to 5), type of establishment (e.g., Hotel, Restaurant, Tourist Attraction), and a specific geographic location among 40 designated Mexican towns classified as "Pueblos Mágicos." These annotations support multiple supervised learning tasks, but in our case, only the polarity label is used as the ground truth reference for evaluating prompt performance.

The reviews span a diverse linguistic landscape, reflecting the informal tone, regional expressions, and varied formatting typical of user-generated content. Such variability presents a significant challenge for automated analysis and provides an ideal setting to test the generalization capabilities of LLMs through prompt-based interaction.

To facilitate a robust evaluation, we filtered and cleaned the dataset to remove malformed or incomplete records. In particular, we excluded entries missing the sentiment label or those with non-standard characters likely to interfere with tokenization. The final corpus retained 207,689 reviews, each with sufficient length and semantic content to be meaningfully interpreted by a generative model.

Figure 1 displays the distribution of sentiment classes across the dataset. As in previous editions of Rest-Mex [10, 12], the data is imbalanced, with class 5 (most positive) being overrepresented, and class 1 (most negative) relatively rare. This imbalance underscores the need for evaluation metrics beyond accuracy, such as macro-averaged F1-scores.

While prior work has leveraged this dataset for fine-tuned classification using Transformers such as BETO [3], we treat it as a static evaluation corpus. No model is trained on the data. Instead, we use the reviews as input queries to evaluate the response quality of pre-trained LLMs under different prompting conditions.

3.2. Prompt Design and Strategies

Prompt engineering requires careful consideration of linguistic structure, specificity, and clarity. In our study, we crafted several prompt templates aimed at eliciting sentiment labels from the model. We designed prompts to simulate typical human instructions, ranging from minimal (zero-shot) to more guided (few-shot) forms. The prompt strategies fall into three categories:

- **Zero-shot Instruction:** The model is asked directly to classify sentiment from the review text, using prompts like *"Classify the sentiment of the following review on a scale from 1 (very negative) to 5 (very positive):"*
- **Few-shot Examples:** The prompt includes one to three labeled examples before the test input. This provides the model with demonstrations to infer the desired task structure and response format.
- **Contextual Cueing:** Some prompts include background information about the "Pueblos Mágicos" initiative or about tourism services, under the hypothesis that contextual framing may improve the model's interpretability of subtle cues.

All prompts are evaluated using the same test samples from the cleaned Rest-Mex dataset. The sentiment predicted by the model (typically a numerical output or a word token like "positive") is post-processed and mapped to the 1–5 polarity scale for metric comparison.

3.3. Model Selection and Query Protocol

We tested our prompts across three major families of commercial and open-source LLMs, including:

- GPT-3.5 and GPT-4 (`text-davinci-003`, `gpt-4`) via the OpenAI API.
- LLaMA-based open models (e.g., Alpaca, Vicuna) deployed locally using HuggingFace pipelines.
- Google PaLM2 through the MakerSuite interface.

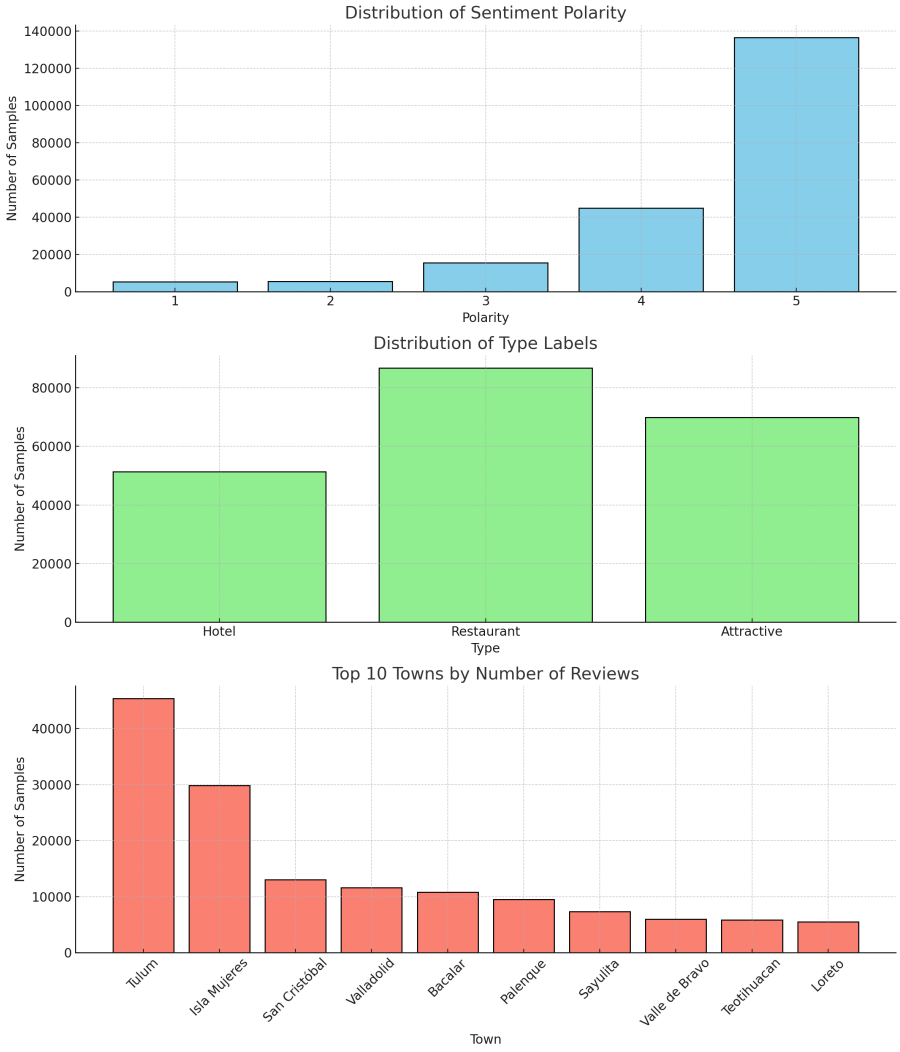

Figure 1: Distribution of class labels for the three tasks: sentiment polarity (top), service type (middle), and town classification (bottom). The polarity distribution is heavily skewed toward positive reviews, type labels favor Hotels and Restaurants, and town labels show significant imbalance.

All models were queried using a fixed protocol: same temperature (0.0 for deterministic outputs), same prompt phrasing (for comparison across models), and a cap of 100 tokens per response. For fairness, the review text was truncated at 500 tokens when necessary to remain within context window limits.

To reduce noise, each prompt-review pair was evaluated three times (where applicable), and the modal predicted label was used for scoring. We also logged model confidence (when provided), latency, and token usage to assess computational cost.

3.4. Evaluation Metrics

The primary metric for model performance is the macro-averaged F1-score, which accounts for class imbalance and reflects balanced precision-recall across sentiment levels. We also report overall accuracy for comparison with existing fine-tuned models [17]. In addition, we performed qualitative error analysis, focusing on reviews that yielded divergent predictions under different prompt types.

Our methodology is designed to isolate the impact of prompt structure on model behavior, independent of parameter tuning or domain-specific retraining. This allows us to evaluate the practical potential of prompt engineering as a lightweight alternative for multilingual sentiment classification in real-world tourism applications.

Table 1
Official evaluation metrics on the Rest-Mex 2025 test set

Task	Macro F1-score	Accuracy
Sentiment Polarity (1–5)	0.199	0.134
Type Classification (3 classes)	0.333	0.825
Town Classification (40 towns)	0.025	0.0087

4. Results

The evaluation of our prompt engineering approach was conducted in two phases: (1) an internal development phase using the training dataset, and (2) an official external evaluation phase using the hidden test set provided by the organizers of Rest-Mex 2025. Given that our methodology did not involve supervised fine-tuning, performance on the training set was expectedly modest, particularly for sentiment polarity.

4.1. Training Set Evaluation (Internal Validation)

In the internal validation phase, prompts were applied to a stratified subset of 10,000 randomly selected reviews from the training corpus. The macro F1-score for sentiment polarity was low across most models, not exceeding 0.25. In particular, reviews with ambiguous or subtle emotional cues—such as sarcastic comments or reviews with mixed opinions—posed significant difficulties for the models. The few-shot prompting variant marginally improved stability but did not yield statistically significant gains in accuracy or F1.

This performance gap on the training set underscores the limitations of prompt-only strategies when dealing with noisy, highly subjective, or domain-specific sentiment tasks in Spanish. It also confirms the hypothesis that prompt engineering, while efficient, may require more specialized design to match the effectiveness of supervised methods in low-resource settings.

4.2. Test Set Evaluation (Official Results)

The official evaluation on the hidden Rest-Mex 2025 test set yielded a mixed but informative performance profile. The results, summarized in Table 1, demonstrate that prompt engineering can provide meaningful signals in certain subtasks, while still struggling in others—particularly when fine-grained differentiation is needed.

Polarity

For polarity classification, our prompt-based approach achieved a macro F1-score of 0.199 and an accuracy of 13.4%. While this is significantly lower than scores typically obtained via fine-tuned models (which often surpass 0.55 in F1), it is notable that the model was able to distinguish some degree of polarity gradient without any parameter tuning. Precision and recall were highest for the most frequent class (label 5), reflecting the class imbalance in the dataset.

Type

In contrast, performance on the type classification task was significantly better, with a macro F1-score of 0.333 and an accuracy of 82.5%. This result suggests that prompt-based reasoning is more effective when the semantic distinctions between classes are clear and supported by consistent lexical patterns—such as references to food, lodging, or attractions. These findings are aligned with previous studies indicating that LLMs perform better in classification tasks with discrete and orthogonal categories [20].

Town

The most challenging task by far was the identification of the correct “Pueblo Mágico” referenced in each review. Here, the model achieved a macro F1-score of just 0.025 and a classification accuracy below 1%. Given the large number of classes (40) and the often implicit nature of geographic references in the text, this result is not surprising. It highlights the difficulty of location grounding in LLMs without access to external knowledge sources or contextual signals like geotagging.

4.3. Error Patterns and Observations

Qualitative analysis of the model’s responses revealed a few recurring trends. First, in the absence of clear polarity markers, the models often defaulted to the majority class ("very positive"), which inflated recall but suppressed precision. Second, the type classification task benefited from prompt templates that explicitly named category examples ("e.g., Hotel, Restaurant, Attraction"), suggesting that concrete framing aids LLM understanding. Third, the town task frequently failed due to hallucination—models generated plausible but incorrect town names not present in the label space.

Overall, while the results remain behind state-of-the-art supervised approaches, they validate the core potential of prompt engineering as a flexible and lightweight alternative for certain subtasks. With further refinement—such as prompt chaining, task-specific calibration, or use of external geographic knowledge bases—performance in low-resource classification tasks could be substantially improved.

5. Conclusion

This study explored the application of prompt engineering techniques for sentiment analysis in the tourism domain, using Spanish-language reviews from the *Rest-Mex 2025* dataset. By leveraging large language models (LLMs) through zero-shot and few-shot prompting, we aimed to assess whether such models could infer sentiment polarity and other attributes without the need for fine-tuning or domain-specific retraining.

Our results reveal a nuanced picture. While prompt-based methods yielded promising performance in the classification of establishment type—achieving an accuracy above 82%—their effectiveness was considerably lower in tasks requiring deeper contextual reasoning, such as sentiment polarity ($F1_{macro}$ = 0.199) and especially town identification ($F1_{macro}$ = 0.025). These outcomes suggest that LLMs, when guided solely by prompts, can capture surface-level patterns and categorical cues, but struggle with fine-grained sentiment interpretation and geospatial grounding, particularly in resource-constrained linguistic environments.

Importantly, our findings validate the core potential of prompt engineering as a viable approach for rapid prototyping, especially in low-resource settings where labeled data or computational power is limited. Unlike traditional supervised methods, prompt-based approaches require no parameter updates and can adapt quickly across domains. However, the tradeoff lies in precision, control, and interpretability.

Future work should focus on improving prompt strategies through dynamic chaining, task-aware calibration, and the integration of external knowledge sources such as tourism gazetteers or sentiment lexicons. Additionally, a deeper exploration of linguistic and cultural factors in prompt design may help unlock more consistent performance in subjective tasks like sentiment analysis.

In sum, while prompt engineering is not yet a substitute for supervised fine-tuning in complex NLP tasks, it represents a compelling and accessible frontier—especially for domain-specific applications like tourism analytics in Spanish-speaking contexts.

Declaration on Generative AI

We declare that the present manuscript has been written entirely by the authors and that no generative artificial intelligence tools were used in its preparation, drafting, or editing.

References

[1] R. Guerrero-Rodriguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, Current issues in tourism 26 (2023) 289–304.

[2] A. Diaz-Pacheco, M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence 36 (2024) 1415–1445.

[3] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Á. Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of king Saud university-computer and information sciences 34 (2022) 10125–10144.

[4] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities 10 (2024) 639–661.

[5] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-González, R. Aranda, Quantifying differences between ugc and dmo's image content on instagram using deep learning, Information Technology & Tourism 26 (2024) 293–329.

[6] E. P. Ramirez-Villaseñor, H. Pérez-Espinosa, M. A. Álvarez-Carmona, R. Aranda, Design, development, and evaluation of a chatbot for hospitality services assistance in spanish, Acta universitaria 33 (2023).

[7] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: a study case in mexico, in: Mexican international conference on artificial intelligence, Springer, 2021, pp. 184–195.

[8] A. Diaz-Pacheco, M. A. Álvarez-Carmona, A. Y. Rodríguez-González, H. Carlos, R. Aranda, Measuring the difference between pictures from controlled and uncontrolled sources to promote a destination. a deep learning approach (2023).

[9] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism (2021).

[10] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022) 289–299.

[11] M. Á. Alvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022).

[12] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, V. Muñiz-Sánchez, A. P. López-Monroy, F. Sánchez-Vega, L. Bustio-Martínez, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023) 425–436.

[13] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Herrera-Semenets, Overview of rest-mex at iberlef 2025: Researching sentiment evaluation in text for mexican magical towns, volume 75, 2025.

[14] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[15] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science 50 (2024) 568–589.

[16] I. Castillo-Ortiz, M. Á. Álvarez-Carmona, R. Aranda, Á. Díaz-Pacheco, Evaluating culinary skill transfer: A deep learning approach to comparing student and chef dishes using image analysis, International Journal of Gastronomy and Food Science 38 (2024) 101070.

[17] M. Á. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, Computación y Sistemas 26 (2022) 977–987.

[18] Y. Zhang, R. Yang, X. Xu, R. Li, J. Xiao, J. Shen, J. Han, Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 2032–2042.

[19] M. Sallam, K. Al-Mahzoum, R. A. Almutawaa, J. A. Alhashash, R. A. Dashti, D. R. AlSafy, R. A. Almutairi, M. Barakat, The performance of openai chatgpt-4 and google gemini in virology multiple-choice questions: a comparative analysis of english and arabic responses, BMC Research Notes 17 (2024) 247.

[20] V. G. Morales-Murillo, H. Gómez-Adorno, D. Pinto, I. A. Cortés-Miranda, P. Delice, Lke-iimas team at rest-mex 2023: Sentiment analysis on mexican tourism reviews using transformer-based domain adaptation (2023).