Multitask Analysis of Spanish Travel Reviews: Sentiment, Destination, and Topic Classification with RoBERTa and LLaMA Ensembles

Carlos Minutti-Martinez[1,*], Boris Escalante-Ramirez[2] and Jimena Olveres-Montiel[2]

[1]*INFOTEC, Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, Mexico*

[2]*Departamento de Procesamiento de Señales, División de Ingeniería Eléctrica, Universidad Nacional Autónoma de México, Mexico City, Mexico*

Abstract

The recent advancements in Large Language Models (LLMs) have significantly improved sentiment analysis by enabling better understanding of nuanced emotions and contextual information. This has led to their widespread adoption across various domains, including healthcare, finance, social media monitoring, market research, and public policy. In the tourism sector, sentiment analysis of visitor reviews plays a key role in understanding tourist satisfaction, influencing travel decisions, improving service quality, supporting reputation management, and enabling personalized travel recommendations. Despite the development of fine-tuned models across different languages, most LLMs are originally trained on English corpora, potentially affecting their performance in other languages such as Spanish. Therefore, evaluating their effectiveness in Spanish-language contexts is essential. In this work, we present the system developed by team AxoloTux for the Rest-Mex 2025 challenge (Researching Sentiment Evaluation in Text for Mexican Magical Towns) at IberLEF 2025. The task involves analyzing Spanish-language reviews of Mexican tourist destinations to predict sentiment polarity, classify the type of attraction, and identify the associated Magical Town. Our approach involves fine-tuning multiple RoBERTa variants, along with LLaMA 3.2 models (1B and 3B parameters), and combining them using an ensemble strategy to enhance robustness. Models were trained and validated on 70% of the dataset, while the remaining 30% was held out as a test set. Ensemble weights were determined based on validation performance. This strategy achieved a final track score of 0.7226 on the test set, securing second place in the competition, closely trailing the top score of 0.7254.

Keywords

Sentiment Analysis, Large Language Models, Spanish NLP, RoBERTa, LLaMA, Model Ensembling, Tourism Reviews, Text Classification, Domain Adaptation

1. Introduction

Sentiment analysis has become a valuable tool in a wide range of domains, including e-commerce, tourism, social media monitoring, healthcare, education, and finance. By extracting subjective information from user-generated content, organizations can better understand customer satisfaction, monitor brand reputation, adapt services, and make data-driven decisions. In the tourism and hospitality industry, for example, analyzing tourist reviews allows stakeholders to assess service quality, identify strengths and weaknesses, and improve the overall visitor experience [1, 2, 3, 4].

Despite its growing relevance, sentiment analysis faces numerous challenges. Models often struggle with ambiguous or sarcastic language, context-dependent expressions, and informal tones. Domain adaptation remains an issue, as models trained in one domain (e.g., product reviews) may perform poorly in others (e.g., travel narratives). Multilingualism and code switching, particularly common in regions with high linguistic diversity, further complicate the analysis. Furthermore, user-generated

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

carlos.minutti@infotec.mx (C. Minutti-Martinez); boris@cecav.unam.mx (B. Escalante-Ramirez); jolveres@cecav.unam.mx (J. Olveres-Montiel)

https://cminuttim.github.io (C. Minutti-Martinez); https://lapi.fi-p.unam.mx/ (B. Escalante-Ramirez); https://lapi.fi-p.unam.mx/ (J. Olveres-Montiel)

0000-0002-4002-0773 (C. Minutti-Martinez); 0000-0003-4936-8714 (B. Escalante-Ramirez); 0000-0002-1514-4520 (J. Olveres-Montiel)

content is frequently noisy and unstructured, containing slang, abbreviations, emojis, and grammatical inconsistencies that hinder pre-processing and accurate sentiment extraction [1, 3, 5].

These challenges are particularly pronounced in Spanish-language texts. Compared to English, there are fewer annotated datasets and linguistic resources available for Spanish, which limits the development of robust models. In addition, the language's complex grammar, regional variations, and idiomatic expressions add another layer of difficulty in generalizing across contexts. Tourist reviews in Spanish often include mixed emotions, informal phrasing, and even code-switching between Spanish and English, requiring models capable of handling nuanced and multilingual inputs [6, 7, 8, 9].

Before the rise of transformer-based models, common approaches for sentiment and text classification included Bag-of-Words (BoW) and TF-IDF representations combined with traditional machine learning classifiers such as Logistic Regression, Support Vector Machines (SVM), or Naive Bayes. Word embedding models such as Word2Vec, GloVe, or FastText were often used alongside neural networks or traditional classifiers. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, were used to capture sequential context but struggled with long-range dependencies. Convolutional Neural Networks (CNNs) were also applied to sequences of word embeddings for feature extraction. Transformer-based architectures overcame many of these limitations by eliminating recurrence and avoiding vanishing / explosive gradient problems, leading to faster and more efficient training [10, 11].

Models such as BERT [12], RoBERTa [13], and XLNet [14] have become standard tools in natural language processing (NLP) for tasks such as sentiment analysis and text classification. These models are pre-trained on massive corpora and fine-tuned for specific downstream tasks. Empirical evidence shows that transformer-based models consistently outperform earlier methods [15, 16, 10]. In recent years, Large Language Models (LLMs) such as LLaMA [17] have demonstrated superior performance on text classification tasks compared to Small Language Models (SLMs) such as BERT and RoBERTa. However, LLMs introduce significantly higher computational costs despite their effectiveness [18].

In this context, the **Rest-Mex 2025**[19] challenge (Researching Sentiment Evaluation in Text for Mexican Magical Towns), part of **IberLEF 2025**[20], presents a relevant and timely benchmark. The goal is to analyze Spanish-language reviews of tourist destinations in Mexico and classify them along three dimensions: sentiment polarity, type of site, and associated *Pueblo Mágico*. Each review reflects a traveler's experience, typically drawn from platforms like TripAdvisor, and includes rich textual data, metadata, and ratings.

Specifically, the task involves:

- **Sentiment polarity classification**: assigning a score from 1 (very negative) to 5 (very positive) based on the user's original rating.
- **Type-of-site classification**: identifying whether the review refers to a hotel, restaurant, or attraction, using contextual clues and metadata.
- **Destination identification**: determining which of the 40 *Pueblos Mágicos* is being referenced, based on location metadata and review content.

The dataset contains over 200,000 reviews, with 70% designated for training and validation, and the remaining 30% held out for testing. This shared task offers a valuable opportunity to explore multilingual sentiment analysis in a real-world, domain-specific, and culturally rich setting.

In this work, we describe our approach to the Rest-Mex 2025 challenge, which centers on fine-tuning multiple transformer-based models for multilingual sentiment and text classification. Our methodology includes the use of various RoBERTa variants and LLaMA 3.2 models (with 1B and 3B parameters), each trained on different subsets of the data. To enhance generalization and robustness, we adopt an ensemble strategy that aggregates predictions from these diverse models. This approach achieved a final track score of 0.7226 on the held-out test set (comprising 30% of the dataset), securing second place in the competition, closely behind the top score of 0.7254.

2. Methodology

2.1. Overview

From the State-of-the-Art analysis, transformer-based models have consistently outperformed traditional methods in text classification tasks [15, 16, 10]. In particular, the winning team of the previous edition of the Rest-Mex challenge adopted a RoBERTa-based approach [9, 21]. Furthermore, LLMs have shown superior performance compared to SLMs in classification benchmarks [17]. Based on these findings, our methodology leverages the RoBERTa and LLaMA models to fine-tune solutions for the three classification tasks: polarity, type, and town.

2.2. Data

The dataset consists of 208,051 TripAdvisor reviews, corresponding to 70% of the entire corpus (the remaining 30% is reserved as a test set). Each record includes the following fields:

- **Title:** Title of the review (Text).
- **Review:** The full text of the review (Text).
- **Polarity:** Sentiment label from 1 to 5.
- **Town:** Name of the town, with 40 different values (Text).
- **Region:** State where the town is located (Text; auxiliary information).
- **Type:** Place type being reviewed (Hotel, Restaurant, Attractive).

2.2.1. Dataset Statistics

The dataset consists of user-generated reviews in Spanish, covering three main place types: restaurants, hotels, and attractions. Each entry includes a textual review, a numerical sentiment score (Table 1) , a destination town (Table 3), and the type of establishment (Table 2). The data is imbalanced across sentiment classes and towns, with a higher concentration of reviews in popular tourist destinations. Table 4 presents the number of tokens per review (title + review), with a median value of 66 tokens. Only 0. 17% has more than 512 tokens and 0. 05% has more than 768 tokens.

Table 1
Polarity distribution

Class	Instances
1 (Very Bad)	5,441
2 (Bad)	5,496
3 (Neutral)	15,519
4 (Good)	45,034
5 (Very Good)	136,561
Total	**208,051**

Table 2
Type of establishment distribution

Class	Instances
Hotel	51,410
Restaurant	86,720
Attractive	69,921
Total	**208,051**

Table 3
Top and bottom 5 towns by number of reviews

Rank	Town	Count
1	Tulum	45,345
2	Isla Mujeres	29,826
3	San Cristóbal de las Casas	13,060
4	Valladolid	11,637
5	Bacalar	10,822
36	Dolores Hidalgo	909
37	Coatepec	818
38	Cuatro Ciénegas	788
39	Real de Catorce	760
40	Tapalpa	725

Table 4
Summary statistics for the number of tokens per review

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Tokens	13	47	66	86.34	105	1814

2.2.2. Data Imbalance

Despite the observed imbalance, following [21], we did not apply aggressive rebalancing techniques due to their limited impact. A basic Gradient Boosting Machine (GBM) using inverse class weights and review length as features was tested but yielded no improvement. Thus, no further imbalance correction was performed.

2.2.3. Text Normalization

Encoding inconsistencies were addressed using the `ftfy` [22] and `unicodedata` Python libraries to standardize all text inputs to UTF-8.

2.3. Evaluation Metrics

Model performance is evaluated using macro-averaged F1-score for each subtask. This metric gives equal weight to all classes, regardless of their frequency.

Let k denote a model, and $F_i(k)$ its F1-score on class i. We define the following metrics:

Polarity Score (Res$_P$)

The polarity task includes 5 sentiment classes (from 1: Very Bad to 5: Very Good). The score is computed as:

$$\mathrm{Res}_P(k) = \frac{1}{5}\sum_{i=1}^{5} F_i(k) \tag{1}$$

Type Score (Res$_T$)

For classifying place types (Attractive, Hotel, Restaurant), the score is:

$$\mathrm{Res}_T(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \tag{2}$$

Town Score (Res$_{MT}$)

This score is averaged over the 10 most frequent towns in the dataset:

$$\text{Res}_{MT}(k) = \frac{1}{40} \sum_{i=1}^{40} F_{\text{MTL}_i}(k) \tag{3}$$

Overall Score (TrackScore)

The final ranking metric is a weighted average across the three tasks:

$$\text{TrackScore}(k) = \frac{2 \cdot \text{Res}_P(k) + \text{Res}_T(k) + 3 \cdot \text{Res}_{MT}(k)}{6} \tag{4}$$

This weighting emphasizes sentiment and town classification.

2.4. Models

We implemented model ensembles, following the evidence that they improve robustness and accuracy across noisy and heterogeneous data [23, 24, 25]. The selected transformer models included:

RoBERTa and BERT variants:

- `UMUTeam/roberta-spanish-sentiment-analysis` [26]
- `PlanTL-GOB-ES/roberta-large-bne-massive` [27]
- `PlanTL-GOB-ES/roberta-base-bne` [28]
- `edumunozsala/roberta_bne_sentiment_analysis_es` [29]
- `dccuchile/bert-base-spanish-wwm-cased` [30]
- `dccuchile/bert-base-spanish-wwm-uncased` [30]
- `dccuchile/roberta-large-bne-finetuned-qa-mlqa` [31]
- `FacebookAI/xlm-roberta-large` [32]
- `joeddav/xlm-roberta-large-xnli` [33]

LLaMA models:

- `nztinversive/llama3.2-1b-Uncensored` [34]
- `meta-llama/Llama-3.2-3B` [35]

Each model had three separate classification heads (linear layers), one for each task. Hidden layer dimensions were adapted: 768 (RoBERTa-base), 1024 (RoBERTa-large), 2048 (LLaMA 1B), and 3072 (LLaMA 3B). For the maximum input text length, RoBERTa models support up to 512 tokens; for LLaMA, 512 tokens was the default option tested, in addition 768 tokens were also tested for LLaMA 3B.

Input strategies:

1. Concatenation: Title and Review joined with a hyphen.
2. Dual-sentence: Title and Review treated as separate inputs.

Ensemble strategy: For $n \in \{3, 5, 7, 12\}$ top-scoring models, we used weighted voting, where each model's vote was scaled by its F1-score.

Table 5
RoBERTa hyperparameters

Parameter	Value
Batch size	32
Epochs	3 (up to 6 with early stopping)
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup steps	500
FP16	True

Table 6
LLaMA + LoRA hyperparameters

Parameter	Value
Batch size	16
Epochs	3
Learning rate	2×10^{-4}
Weight decay	0.01
LoRA Rank	16
LoRA Alpha	32
LoRA Dropout	0.05
Quantization	NormalFloat 4-bit (NF4)

2.5. Training and Hyperparameters

The most relevant parameters for the RoBERTa and LLaMA models are presented in Tables 5-6.

Training was conducted in two stages. In the first stage, all parameters of the base RoBERTa model were frozen and only task-specific classification heads were trained. This allowed the model to adapt the classifiers to the pre-trained representations without altering the core language model. In the second stage, all parameters were unfrozen, and the entire architecture was fine-tuned jointly.

LLaMA models were quantized using 4-bit precision with `BitsAndBytes` [36], and fine-tuned using LoRA [37] through the PEFT library.

The final predictions were obtained using *weighted voting*, a strategy that combines elements of hard and soft voting. Each model casts a discrete vote for a class, but the influence of its vote is weighted according to its macro-F1 score on the validation set (10% of the training data), thereby giving more importance to more reliable models. Specifically, for each model i, classification task, and record in the test dataset, the predicted class receives a voting weight defined as $w_i = (2 \cdot \mathrm{F1}_{\text{Polarity}} + \mathrm{F1}_{\text{Town}})/3$, while all other classes receive a weight of zero. The final predicted class corresponds to the one with the highest cumulative voting weight across all models.

The *Type* task was excluded from the voting formula because its F1 scores were consistently very high and similar across models, typically close to 1. Including it would have diluted the influence of the other tasks without contributing to meaningful discrimination, so its weight was set to zero.

A workflow diagram of the proposed system is shown in Figure 1.

3. Results

3.1. Validation Results

Table 7 presents the F1 macro scores for the Polarity and Town tasks on the validation split for each model. The number of parameters is reported in billions. For the LLaMA models, the number of parameters trained using the LoRA approach is 24.3 and 11.3 million for the 3B and 1B models, respectively. The *Sentence* column indicates whether the title and review were treated as separate sentences (with a special

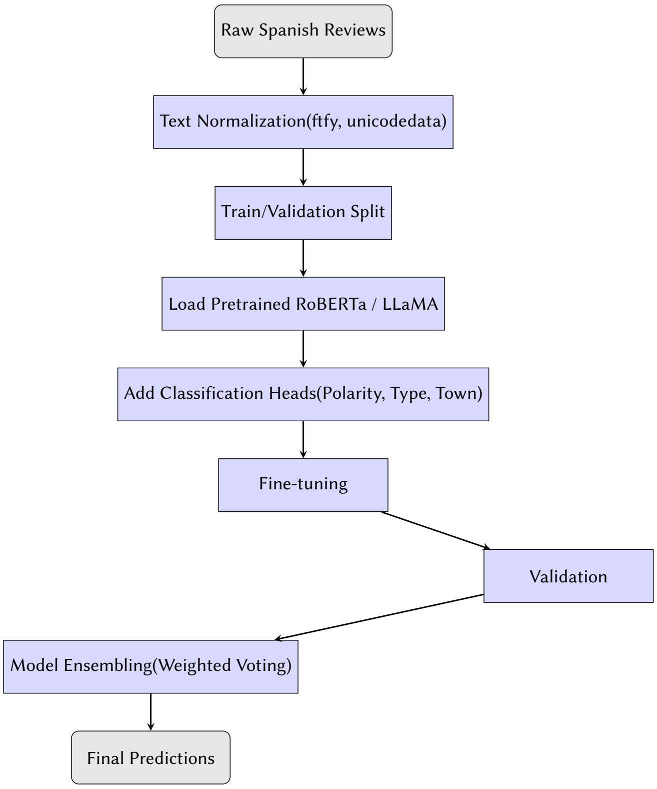
Raw Spanish Reviews
Text Normalization(ftfy, unicodedata)
Train/Validation Split
Load Pretrained RoBERTa / LLaMA
Add Classification Heads(Polarity, Type, Town)
Fine-tuning
Validation
Model Ensembling(Weighted Voting)
Final Predictions

Figure 1: Workflow of the system

token) or concatenated. F1 macro scores are shown for both the Polarity, Type and Town classification tasks. The weighted average combines Polarity and Town tasks using a $2 : 1$ ratio, giving more weight to the polarity classification due to its higher difficulty. The establishment type classification was excluded from this comparison, as all models achieved F1 scores between 0.97 and 0.98. Figure 2 shows a comparison of the F1 scores for each model across the three classification tasks,

Table 7
Validation performance for each model. Weighted average is calculated as $(2 \times \text{F1-Polarity} + \text{F1-Town})/3$.

Rank	Model	Params (B)	Seed	Sentence	Polarity	Type	Town	W.Avg.
0	Llama-3.2-3B (768 Tokens)	3.24	33	YES	**0.6414**	0.9871	0.6777	**0.6535**
1	Llama-3.2-3B	3.24	32	YES	0.6353	**0.9877**	**0.6785**	0.6497
2	llama3.2-1b-Uncensored	1.25	42	YES	0.6223	0.9858	0.6666	0.6371
3	xlm-roberta-large-xnli	0.56	41	YES	0.6326	0.9845	0.6371	0.6341
4	bert-base-spanish-wwm-cased	0.11	43	YES	0.6187	0.9812	0.6336	0.6237
5	roberta-large-bne-massive	0.36	43	NO	0.6146	0.9773	0.6230	0.6174
6	xlm-roberta-large	0.56	42	YES	0.6307	0.9824	0.5846	0.6153
7	roberta-spanish-sentiment-analysis	0.12	42	NO	0.6149	0.9740	0.6084	0.6127
8	roberta_bne_sentiment_analysis_es	0.12	43	NO	0.6182	0.9842	0.5915	0.6093
9	bert-base-spanish-wwm-uncased	0.11	44	YES	0.6033	0.9764	0.6140	0.6069
10	roberta-large-bne-massive	0.36	42	YES	0.5948	0.9806	0.6262	0.6053
11	roberta_bne_sentiment_analysis_es	0.12	42	NO	0.6097	0.9748	0.5891	0.6028
12	roberta-base-bne	0.12	42	NO	0.6046	0.9741	0.5787	0.5960

3.2. Test Results

Table 8 presents the results for different ensembles on the test dataset. The *Run* column indicates the ensemble strategy. Prefix "_T" specifies that the model with 768-token input length (rank 0) was included. The number after "E" indicates how many models were used. For example, "Axolotux_E_T3" is the ensemble of ranks 0–2, while "Axolotux_E3" includes ranks 1–3.

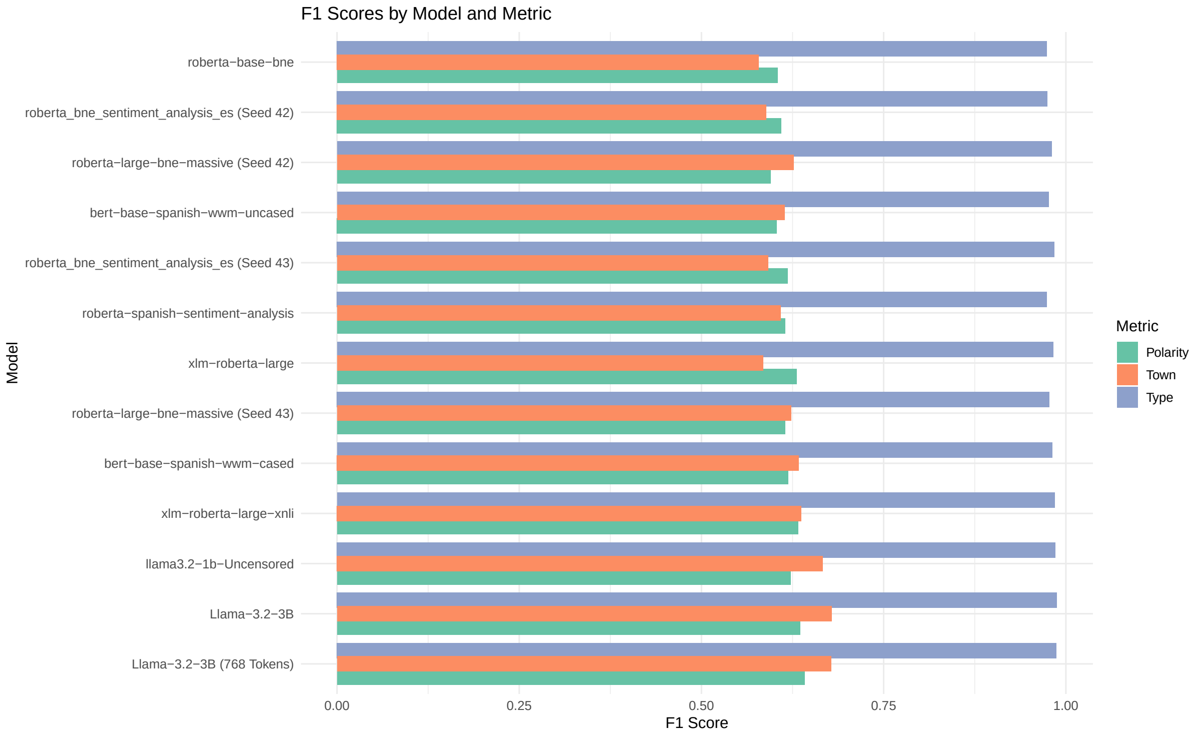

Figure 2: Comparison of F1 scores for each model across three classification tasks: Polarity, Type, and Town. Models are ordered as in Table 7.

Table 8
Test dataset performance for different ensemble strategies.

Run	F1-Polarity	F1-Type	F1-Town	Track Score
Axolotux_E_T3	0.6395	**0.9876**	**0.6896**	**0.7226**
Axolotux_E3	0.6390	0.9874	0.6843	0.7197
Axolotux_E5	0.6403	0.9867	0.6826	0.7192
Axolotux_E7	0.6413	0.9859	0.6751	0.7157
Axolotux_E_T1	0.6333	0.9865	0.6775	0.7143
Axolotux_E1	0.6305	0.9860	0.6782	0.7136
Axolotux_E12	**0.6414**	0.9851	0.6670	0.7115

4. Discussion

From the results in Table 7, the best single model on the validation dataset was `meta-llama/Llama-3.2-3B` using 768 tokens (rank 0), followed closely by its 512-token version (rank 1) and the `llama3.2-1b-Uncensored` model (rank 2). These LLaMA models benefit from a large number of trainable parameters and extensive pretraining.

Among RoBERTa-based models, the best result was achieved by `joeddav/xlm-roberta-large-xnli`, which is fine-tuned on multilingual Natural Language Inference (NLI) data, with the intention of being used for zero-shot classification. Here, it is further fine-tuned for our tasks. Rank 0-3 are multilingual model, suggesting and advantage over language fine-tuned models.

Interestingly, the second-best RoBERTa model (`dccuchile/bert-base-spanish-wwm-cased`) is a smaller model, suggesting that well-pretrained compact models can still provide strong performance for Spanish classification tasks when resources are limited.

Models in which the title and review were processed as separate sentences (with a separator token) slightly outperformed those where the texts were concatenated, suggesting that preserving structural distinctions in the input text can help the model better understand the context.

Table 8 shows the final test scores for different ensemble strategies. The best overall performance was obtained using the top three LLaMA models (Axolotux_E_T3), confirming that combining diverse but strong models leads to improved generalization. The second-best result came from combining ranks 1–3 (Axolotux_E3), and even using five models (Axolotux_E5) produced nearly equivalent performance.

Notably, using a larger ensemble (Axolotux_E12) hurt the performance, which may result from including weaker models that introduce noise. This reinforces the importance of carefully selecting ensemble members. Also, single-model performance (Axolotux_E_T1, Axolotux_E1) was consistently lower than ensemble scores, confirming the benefit of combining predictions.

The polarity classification task is related to that of the 2023 Rest-Mex edition, although the tasks are not exactly the same and only partially share data. The winning model from that year achieved a macro-F1 score of 0.6217 using a single RoBERTa model with domain adaptation [9, 21]. In contrast, our ensemble-based approach achieves a minimum macro-F1 of 0.6379 on this task, suggesting that model diversity and strategic ensembling can outperform domain adaptation alone.

After reviewing the most extreme cases of misclassification (i.e., those with the highest prediction error), we observed that many of these instances appear to be due to noise or mislabeling in the original polarity annotation. In several cases, the model's prediction actually aligns better with the sentiment expressed in the text than the provided ground truth. Below, we present three illustrative examples:

Example 1
Polarity (ground truth): 1 (Very Bad)
Prediction: 5 (Very Good)

> **Original Spanish review:**
> *Sitio sagrado - El hotel es muy bonito, está alejado del pueblo por lo que no hay ruido, si deciden ir al spa hagan la gruta de los sentidos, los masajes son muy buenos; el restaurante tiene una vista espectacular y la comida es rica.*
>
> **English translation:**
> *Sacred site - The hotel is very nice, it is away from the town so there is no noise. If you decide to go to the spa, do the grotto of the senses, the massages are very good; the restaurant has a spectacular view and the food is delicious.*

Example 2
Polarity (ground truth): 5 (Very Good)
Prediction: 1 (Very Bad)

> **Original Spanish review:**
> *Fraude - El hotel no dice por conveniencia de las aguas negras de la playa frente a su hotel. Muy decepcionada porque engañan al turismo. En internet encuentras muchas noticias sobre el tema, la última es de este enero 2022 donde está entre las 5 playas para no visitar.*
>
> **English translation:**
> *Fraud - The hotel does not mention the sewage from the beach in front of it, for convenience. Very disappointed because they deceive tourists. On the internet you find many news reports on the subject; the latest is from January 2022, listing it among the top 5 beaches to avoid.*

Example 3
Polarity (ground truth): 1 (Very Bad)
Prediction: 5 (Very Good)

> **Original Spanish review:**
> *EXCELENTE - un hotel enclavado en la peña, con espectaculares vistas, habitaciones modernas*

> *y lujosas decoradas por diseñadores reconocidos. Amenidades también de diseñador, cojines Swarovski y gran lujo con estilo y excelente gusto. No puedes creer el estar en un lugar tan paradisíaco y con tal confort. Propio para viajeros exigentes y conocedores.*
>
> **English translation:**
> *EXCELLENT - A hotel nestled in the rock, with spectacular views, modern and luxurious rooms decorated by renowned designers. Designer amenities as well, Swarovski cushions and great luxury with style and excellent taste. You can't believe you're in such a paradisiacal place with such comfort. Perfect for discerning and knowledgeable travelers.*

In this context, an ensemble of models is particularly useful for mitigating the effects of noise and potential mislabeling in the data. By aggregating the predictions of multiple models through weighted voting, the ensemble reduces the likelihood of overfitting to spurious patterns or annotation errors present in individual examples. This collective decision-making helps to smooth out inconsistencies and emphasizes consensus among more reliable models, which is especially valuable when dealing with subjective or noisy tasks such as sentiment classification.

5. Conclusions

This work evaluated a wide range of transformer-based models for multilingual and Spanish-specific text classification in the context of the Rest-Mex 2025 challenge. The most effective single models were those based on the LLaMA 3.2 architecture, with the 768-token variant outperforming all others. Among RoBERTa models, `xlm-roberta-large-xnli` and `bert-base-spanish-wwm-cased` achieved the best results, offering a viable alternative for low-resource setups.

We also explored different ways of encoding the title and review, finding that treating them as separate input segments marginally improved performance.

Model ensembling provided significant benefits. A weighted voting strategy, based on individual model F1 scores, allowed us to combine the strengths of multiple models. The best ensemble (Axolotux_E_T3) surpassed all single models, demonstrating that even small ensembles (3–5 members) can provide notable gains. However, overly large ensembles (e.g., 12 models) degraded performance, indicating the importance of selective ensembling.

Compared with prior solutions, such as the 2023 Rest-Mex winner based on domain adaptation, our ensemble approach provided superior performance, especially in the polarity classification task. These results suggest that ensembling diverse pre-trained models can be a robust alternative to domain-specific fine-tuning.

Acknowledgments

This research was supported in part by UNAM PAPIIT grants IT101624 and IN108624.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to improve the grammar and clarity of this manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: Approaches, challenges and trends, Knowledge-Based Systems 226 (2021) 107134.

[2] 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 (2023) 639–661. URL: http://dx.doi.org/10.1108/IJTC-09-2022-0223. doi:10.1108/ijtc-09-2022-0223.

[3] S. Sushma, S. K. Nayak, M. V. Krishna, A comprehensive review of sentiment analysis: Trends, challenges, and future directions, in: 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), IEEE, 2024, pp. 1175–1181.

[4] Á. Díaz-Pacheco, R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, A. Y. Rodríguez-GonzÁlez, R. Aranda, A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101746. URL: http://dx.doi.org/10.1016/j.jksuci.2023.101746. doi:10.1016/j.jksuci.2023.101746.

[5] N. Raghunathan, K. Saravanakumar, Challenges and issues in sentiment analysis: A comprehensive survey, IEEE Access 11 (2023) 69626–69642.

[6] 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. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[7] 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).

[8] M. A. 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, Ángel 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. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003615. doi:https://doi.org/10.1016/j.jksuci.2022.10.010.

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

[10] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, L. He, A survey on text classification: From traditional to deep learning, ACM Transactions on Intelligent Systems and Technology (TIST) 13 (2022) 1–41.

[11] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artificial Intelligence Review 55 (2022) 5731–5780.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[15] F. Wang, Comparative evaluation of sentiment analysis methods: From traditional techniques to advanced deep learning models, Applied and Computational Engineering 105 (2024) 23–29.

[16] T. Obinwanne, P. Brandtner, Enhancing sentiment analysis with gpt—a comparison of large language models and traditional machine learning techniques, in: International conference on WorldS4, Springer, 2023, pp. 187–197.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal,

E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[18] W. Cunha, L. Rocha, M. A. Gonçalves, A thorough benchmark of automatic text classification: From traditional approaches to large language models, arXiv preprint arXiv:2504.01930 (2025).

[19] 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.

[20] 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.

[21] 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).

[22] R. Speer, ftfy, Zenodo, 2019. URL: https://doi.org/10.5281/zenodo.2591652. doi:10.5281/zenodo.2591652, version 5.5.

[23] A. Mohammed, R. Kora, An effective ensemble deep learning framework for text classification, Journal of King Saud University-Computer and Information Sciences 34 (2022) 8825–8837.

[24] A. Bari, G. Saatcioglu, Emotion artificial intelligence derived from ensemble learning, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, 2018, pp. 1763–1770.

[25] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, M. Kumar, A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques, Artificial Intelligence Review 56 (2023) 13407–13461.

[26] UMUTeam, roberta-spanish-sentiment-analysis, https://huggingface.co/UMUTeam/roberta-spanish-sentiment-analysis, 2023. Accessed: 2025-06-03.

[27] Text Mining Unit (TeMU), Barcelona Supercomputing Center, roberta-large-bne-massive, https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-massive, 2022. Accessed: 2025-06-03.

[28] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[29] E. M. Sala, roberta_bne_sentiment_analysis_es, https://huggingface.co/edumunozsala/roberta_bne_sentiment_analysis_es, 2022. Accessed: 2025-06-03.

[30] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[31] D. UChile, roberta-large-bne-finetuned-qa-mlqa, https://huggingface.co/dccuchile/roberta-large-bne-finetuned-qa-mlqa, 2023. Accessed: 2025-06-03.

[32] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[33] joeddav, xlm-roberta-large-xnli, https://huggingface.co/joeddav/xlm-roberta-large-xnli, 2023. Accessed: 2025-06-03.

[34] nztinversive, llama3.2-1b-uncensored, https://huggingface.co/nztinversive/llama3.2-1b-Uncensored, 2024. Accessed: 2025-06-03.

[35] Meta, Llama-3.2-3b, https://huggingface.co/meta-llama/Llama-3.2-3B, 2024. Accessed: 2025-06-03.

[36] E. Frantar, S. Ashkboos, T. Hoefler, D.-A. Alistarh, Optq: Accurate post-training quantization for generative pre-trained transformers, in: 11th International Conference on Learning Representations, 2023.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021). URL: https://arxiv.

org/abs/2106.09685.