Tourist Reviews Analysis: An Integral Approach with Traditional Models and Fine-Tuned LLMs

Alberto Esteben Reyes-Peralta[1,2,*], Edgar Abidán Padilla Luis[1], Victor Muñiz Sánchez[2] and David Pinto[1]

[1]Faculty of Computer science, Benemérita Universidad Autónoma de Puebla, Puebla, México.
[2]Centro de Investigación en Matemáticas, Monterrey, México.

Abstract

Sentiment analysis is a key task in Natural Language Processing, involving the computational understanding of opinions and emotions in text. In this work, we participate in the Rest-Mex track at IberLEF 2025, which includes three tasks: sentiment polarity classification (scale 1–5), attraction type identification (Hotel, Restaurant, Attraction), and classification of the *Pueblo Mágico* mentioned in the review. Our methodology follows a unified pipeline of exploratory data analysis, preprocessing, and model training using both traditional and advanced approaches, including fine-tuned large language models. For the polarity task, the fine-tuned transformer model achieved a macro F1-score of 0.6155, demonstrating strong performance on the majority class but reduced effectiveness on minority classes due to class imbalance. The attraction type classification obtained a high F1-score of 0.9761, with excellent precision and recall across all categories. For the *Pueblo Mágico* classification, a KNN classifier using TF-IDF and cosine similarity improved over the baseline with an F1-score of 0.57, although class imbalance limited its generalization capabilities. All proposed models outperformed the official baseline, demonstrating the effectiveness of our approaches. Future work includes applying class balancing techniques and generating synthetic data with large language models to improve minority class representation and overall model robustness.

Keywords

Sentiment Analysis, Natural Language Processing, Rest-Mex Track, IberLEF 2025.

1. Introduction

Sentiment analysis is a core task in Natural Language Processing (NLP) that involves the computational study of opinions, attitudes, and emotions expressed in textual content such as reviews, comments, or social media posts directed at specific entities [1]. In recent years, this area has garnered increasing attention, leading to a wide range of real-world applications across diverse domains, including politics [2], marketing [3, 4], and social media analysis [5, 6, 7]. The growing demand for automated sentiment analysis solutions, coupled with the linguistic complexity and context-dependence of human expression, underscores the need for continued research aimed at enhancing the performance and robustness of sentiment classification techniques. Advancing this line of research is particularly important for underrepresented languages and domain-specific applications, where traditional models often struggle to achieve competitive accuracy.

In this context, the Rest-Mex track [8, 9, 10, 11] at IberLEF 2025 [12] promotes research in opinion mining related to various tourist attractions. This year's edition includes three main tasks: identifying the sentiment polarity of opinion texts on a scale from 1 to 5, classifying the type of destination (Hotel, Restaurant, or Attraction), and identifying the name of the Magical Town (*Pueblo Mágico*) from which the review originates.

In this work, we carry out a wide series of experiments with different text representations, ranging from wrod-frequency representations to fine-tuned large language models (LLM), and different

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
alberto.reyes@alumno.buap.mx (A. E. Reyes-Peralta); edgar.abidan.pl@gmail.com (E. A. P. Luis); victor_m@cimat.mx (V. M. Sánchez); david.pinto@correo.buap.mx (D. Pinto)
0009-0006-7195-5340 (A. E. Reyes-Peralta); 0009-0002-9939-884X (E. A. P. Luis); 0000-0003-4810-3714 (V. M. Sánchez); 0000-0002-8516-5925 (D. Pinto)

supervised and semi-supervised approaches to tackle the three tasks of RestMex.

The paper is organized as follows: in Section 2 we mention briefly some works related to the tasks and our approaches. In Section 3 we describe our proposed approaches for all tasks, including data preprocessing and some exploratory data analysis. In Section 4 we present the main results we obtained and finally, in Section 5 we present our conclusions and future work related to our proposal.

2. Related Work

Several recent studies have explored sentiment analysis using both traditional and deep learning techniques. Abd El-Jawad *et al.* [13] compared machine learning and deep learning algorithms on over one million tweets, proposing a hybrid system based on text mining and neural networks. Their approach achieved an accuracy of 83.7%, surpassing traditional supervised methods.

Kokab *et al.* [14] introduced an enhanced hybrid sentiment analysis model (CBRNN) combining BERT with dilated convolution and Bi-LSTM for sentence-level classification across four domains (airlines, autonomous vehicles, presidential elections, and movies). Initially annotated using zero-shot classification and evaluated via precision, recall, F1-score, and AUC, their model outperformed Glove and Word2Vec embeddings, achieving up to 0.4% improvement in AUC.

Kauffmann *et al.* [3] developed a multilevel sentiment analysis model (document, sentence, and aspect-level) integrating star ratings, prices, and sentiment polarities for marketing decision-making. Validated on Amazon data, their system enhanced product recommendation, brand management, and sustainable consumption strategies.

Mann *et al.* [15] proposed an Enhanced BERT model for Twitter sentiment analysis, achieving 96% accuracy in classifying emotions such as happiness, sadness, and anger, using preprocessing tailored to social media language.

Singh *et al.* [16] applied BERT to analyze COVID-19-related tweets, comparing sentiment in posts from India and the rest of the world. The model reached 94% accuracy, revealing more positive sentiment in Indian tweets and insights into perceptions of governmental actions.

While the use of complex deep learning architectures and LLMs has set the trend in recent years for addressing text analysis and NLP problems in general, basic text representation methods related to bag of words or TF-IDF continue to show remarkable results in several related tasks [17, 18], with the computational advantages this represents. For this reason, one of our aims is to compare state-of-the-art models for text representations and the basic ones together with standard classification algorithms.

3. Methodology

3.1. Exploratory Data Analysis

The distribution of the Polarity variable is summarized by the counts and proportions shown in Fig. 1. The majority of instances are concentrated in the higher polarity categories, with Polarity 5 representing 136,561 instances (65.64%), followed by Polarity 4 with 45,034 instances (21.65%). Lower polarity values occur less frequently, with Polarity 3, 2, and 1 comprising 7.46%, 2.64%, and 2.62% of the data, respectively. Descriptive statistics indicate a mean polarity of 4.45 with a standard deviation of 0.93, a minimum of 1, and a maximum of 5. The 25th percentile is at 4, the median at 5, and the 75th percentile also at 5, reflecting a strong skew toward positive sentiment in the dataset.

The distribution of data by Type is shown in Fig. 2. As observed, Type 0 (Restaurant) is the most frequent, accounting for a total of 86,720 instances, which represents approximately 41.68% of the dataset. It is followed by Type 1 (Attractive) with 69,921 instances (33.61%) and Type 2 (Hotel) with 51,410 instances (24.71%). From a statistical standpoint, the Type variable comprises 208,051 observations, with a mean of 0.830 and a standard deviation of 0.797, indicating a slight concentration toward the lower type values. The minimum value is 0 and the maximum is 2. Additionally, 25% of the instances fall within Type 0, the median (50%) corresponds to Type 1, and 75% of the data also belong to Type 1.

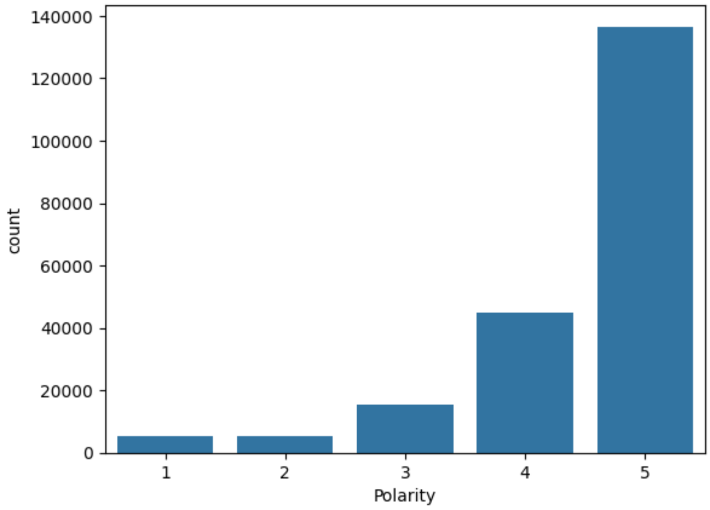

Figure 1: Distribution of review counts by Polarity levels, showing a predominance of positive sentiment with most instances clustered at the highest polarity scores.

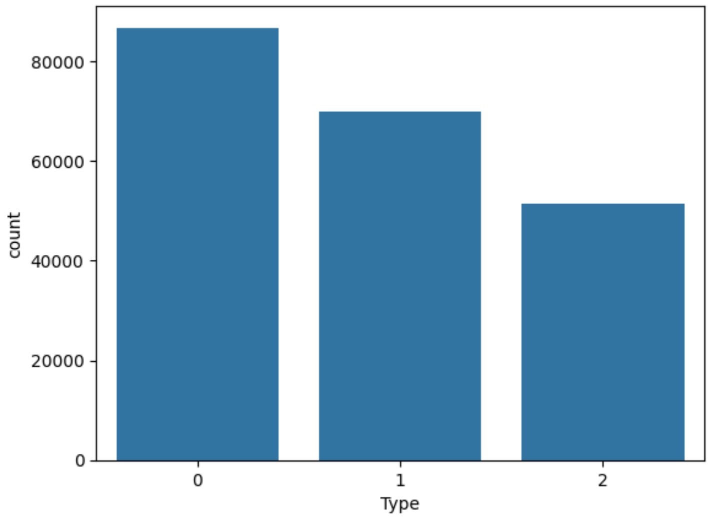

Figure 2: Distribution of instances across Type categories: Restaurant (0), Attractive (1), and Hotel (2), highlighting class imbalance with a higher concentration in categories 0 and 1.

This distribution suggests a non-uniform class balance, with a higher prevalence of instances in Types 0 and 1.

For *Pueblo Mágico* the dataset contains reviews from 40 different towns, with a notable imbalance in their representation, as shown in Fig. 3. The most frequent town is Tulum, contributing 45,345 instances (21.80%), followed by Isla Mujeres with 29,826 instances (14.34%), and San Cristóbal de las Casas with 13,060 instances (6.28%). These three towns alone account for over 40% of the total data. At the other end of the spectrum, towns such as Cuatro Cienegas, Real de Catorce, and Tapalpa have much smaller representations, with 788 (0.38%), 760 (0.37%), and 725 (0.35%) instances respectively. The remaining towns exhibit intermediate frequencies, ranging between these extremes, which highlights a skewed distribution where a few towns dominate the dataset while many others are underrepresented. This imbalance should be taken into account when performing analyses or training models to avoid bias toward the majority locations.

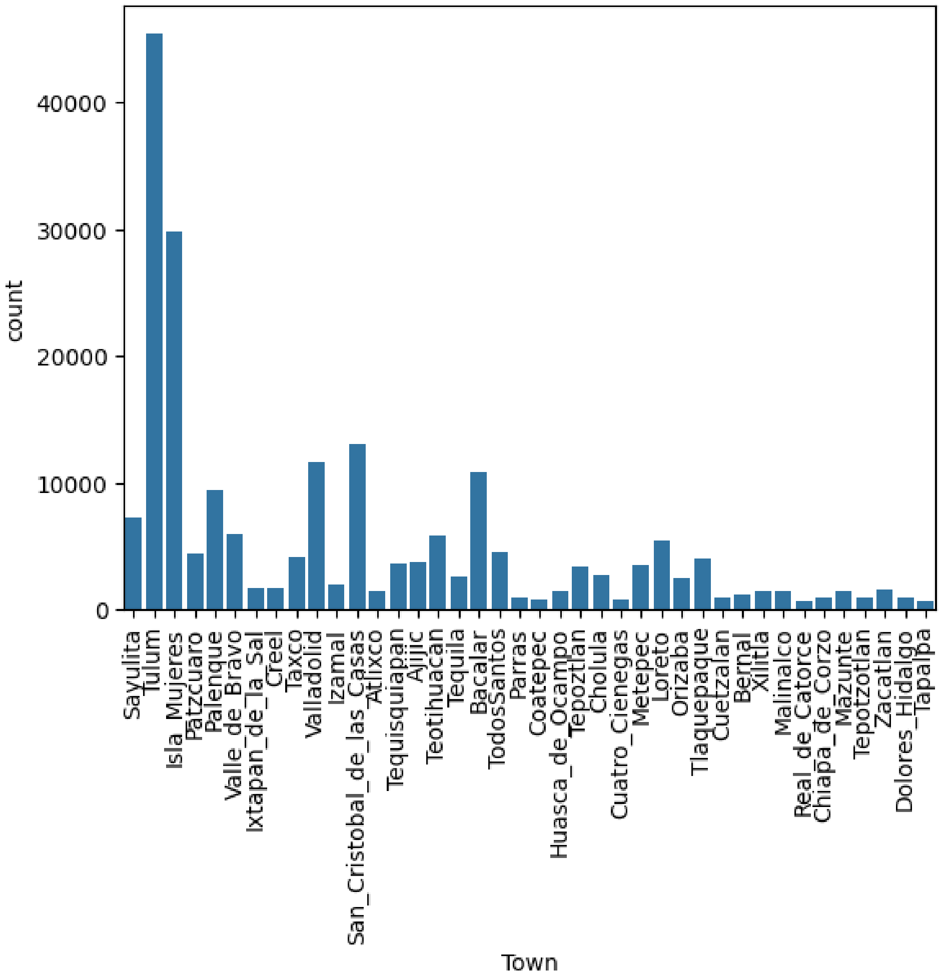

Figure 3: Distribution of review counts across towns, highlighting the concentration in a few major locations such as Tulum and Isla Mujeres, contrasted with many towns with substantially fewer review.

3.2. Data Preprocessing

Data preprocessing is a crucial step in natural language processing, aimed at transforming raw text into a structured format suitable for machine learning and analytical tasks. This process enhances data quality by removing noise and standardizing text, thereby improving model performance and interpretability.

The preprocessing pipeline was applied to a concatenated text field derived from the `Title` and `Review` columns of the dataset. This procedure included the following steps: (1) normalization of accented characters (e.g., á to a); (2) removal of common Spanish stop-words (e.g., `el`, `la`) to reduce lexical noise; (3) lemmatization to obtain canonical word forms (e.g., `comiendo` to `comer`) using a Spanish lemmatizer; (4) elimination of all punctuation marks; and (5) removal of special characters, such as emojis and hashtags.

3.3. Models and Pipelines

The general methodology applied to the three tasks, polarity classification, attraction type identification, and town identification, followed a consistent workflow, as shown in Fig. 4. For each task, exploratory data analysis was conducted to understand the structure and specific characteristics of the corresponding corpus. Subsequently, the data were preprocessed and split into training, validation, and test sets. Various models were evaluated for each task, selecting and fine-tuning those with the best performance using the training and validation sets. Finally, the test set was used for evaluation, and the results were

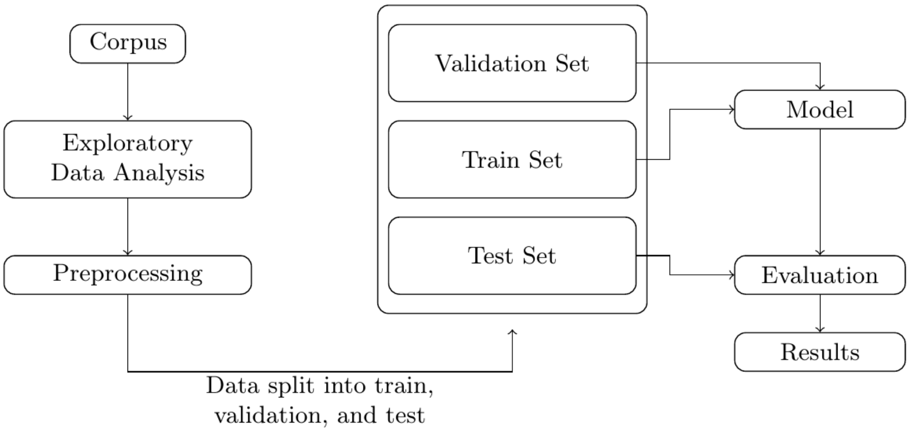

Figure 4: Workflow followed for transformer fine-tuning on the polarity classification task.

compiled to draw relevant conclusions for each problem addressed. The evaluation on the test sets was carried out using standard metrics such as accuracy, precision, recall, and F1-score to assess the effectiveness of the models.

Regarding the polarity and attraction type, after partitioning the data into training, validation, and test sets, fine-tuning was performed using BETO [19], a pre-trained Spanish language model. Specifically, we used the dccuchile/bert-base-spanish-wwm-cased model, which incorporates Whole Word Masking (WWM) and retains case sensitivity features that enhance the model's ability to learn robust semantic and syntactic representations. BETO follows the *BERT base* configuration, comprising 12 Transformer layers, 768 hidden dimensions, 12 attention heads, and approximately 110 million parameters. In our architecture, BETO was used as a frozen encoder to generate contextual embeddings from the `[CLS]` token, summarizing the input sequence. These embeddings were then passed through a feed-forward classifier consisting of two linear layers separated by a ReLU activation function, reducing the dimensionality from 768 to 50 and outputting logits for classification. Fine-tunning was conducted for 2 epochs with a batch size of 16 and a maximum sequence length of 128 tokens. Optimization was performed using the Adam algorithm with a learning rate of `5e-5` and epsilon of `1e-8`. The loss function employed was `CrossEntropyLoss`, appropriate for multi-class classification tasks. The model was trained on the training set and validated on the validation set to optimize its performance. Subsequently, it was evaluated on the test set to assess its generalization capability. Lastly, performance metrics were reported to identify the model's strengths and limitations in each task.

The Magical Town identification task utilized two datasets: one with descriptions of 40 Magical Towns extracted from Spanish Wikipedia using `wikipediaapi` [1], and a training dataset from the Rest-Mex 2025 challenge. Four text representation methods were implemented: initial TF-IDF[20] vectorization of descriptions to create sparse vectors; BETO[21] embeddings reduced via UMAP[22]; RAKE[23] combined with TF-IDF to extract and vectorize key phrases, though limiting semantic context; and TF-IDF with data augmentation, combining Magical Towns descriptions with preprocessed reviews training data to increase class instances. Initial KNN classifiers (TF-IDF, BETO with UMAP, RAKE with TF-IDF) were tested with k=1–15 and various distance metrics. When using representations based on Wikipedia descriptions, we approach this task as a semi-supervised problem, since our goal is to identify which description is closest (in the embedding space) to each opinion about the magical town, and assign the location based on this criterion. This is why we use a simple KNN classifier. We were able to observe that, augmenting the dataset with reviews from training data, improved representativeness. A final KNN classifier with TF-IDF, optimized at k=15 with cosine metric, was trained on the augmented dataset and saved for evaluation on the reviews test set, achieving improved accuracy and robustness.

[1] https://github.com/martin-majlis/Wikipedia-API

4. Results

4.1. Task 1: Polarity Analysis

The performance of the fine-tuned BETO model was evaluated on the test set, which maintains the original class distribution and includes samples not seen during training. Table 1 presents the precision, recall, and F1-score for each class. The majority class (class 4) achieved the best results, with a precision of 0.8486, recall of 0.9127, and F1-score of 0.8795, reflecting its overrepresentation in the dataset. In contrast, minority classes (classes 0 and 1) obtained considerably lower scores, underscoring the model's limitations in handling infrequent categories. Overall, the model reached an accuracy of 76.84% and a macro F1-score of 0.6155, indicating acceptable performance but revealing room for improvement, particularly in addressing class imbalance.

Class	Precision	Recall	F1-score
0	0.7129	0.6710	0.6913
1	0.4425	0.4127	0.4271
2	0.6120	0.5316	0.5690
3	0.5625	0.4677	0.5107
4	0.8486	0.9127	0.8795
Macro Average	0.6357	0.5991	0.6155
Weighted Average	0.7547	0.7684	0.7596
Accuracy		0.7684	

Table 1: Classification report of the fine-tuned model on the test set for the polarity analysis task.

The confusion matrix shown in Fig. 5 reveals that the model performs notably well on class 4, indicating a strong bias toward this overrepresented category. In contrast, considerable misclassification is observed in the lower-polarity classes, particularly classes 0 and 1, which are frequently confused with other labels. Furthermore, the model tends to misclassify samples between adjacent classes—such as 2 and 3, or 3 and 4, suggesting limited sensitivity to subtle distinctions in sentiment polarity. These issues are primarily attributed to the pronounced class imbalance in the dataset, which favors majority classes and impairs the model's ability to generalize across underrepresented categories.

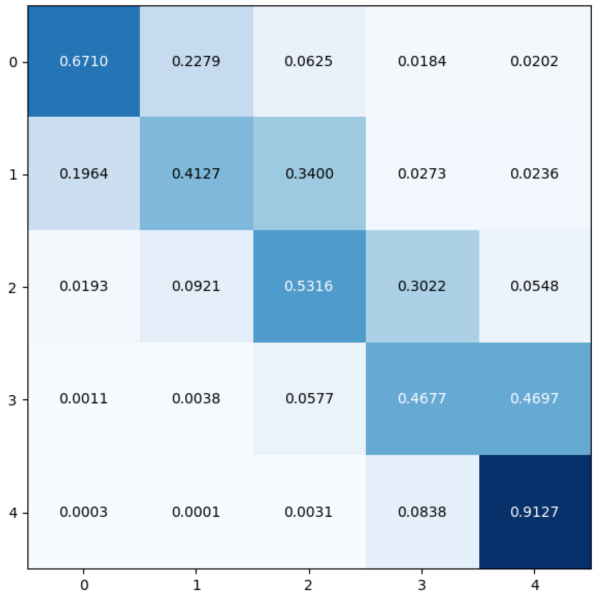

Figure 5: Normalized confusion matrix for polarity classification..

4.2. Task 2: Atraction Identification

The evaluation metrics for the fine-tuned BETO model on the test set are summarized in Table 2. The model achieved high precision, recall, and F1-score across all three classes, with overall accuracy reaching 97.61%. Type 0 (Restaurant) obtained the highest F1-score of 0.9809, followed closely by Type 1 (Attractive) with 0.9762, and Type 2 (Hotel) with 0.9679. The weighted averages indicate a balanced performance, confirming the model's effectiveness in the multi-class classification task.

Type	Precision	Recall	F1-score	Support
0 (Restaurant)	0.9794	0.9824	0.9809	8,672
1 (Attractive)	0.9764	0.9760	0.9762	6,993
2 (Hotel)	0.9701	0.9658	0.9679	5,141
Accuracy		0.9761		
Macro avg	0.9753	0.9747	0.9750	20,806
Weighted avg	0.9761	0.9761	0.9761	20,806

Table 2: Classification report of the fine-tuned model on the test set for the attraction type identification task.

The analysis of the confusion matrix (see Fig. 6) reveals that the model achieves high precision across all attraction categories, with a strong concentration of correct predictions along the main diagonal. Misclassifications are relatively infrequent and uniformly distributed, indicating that the model does not exhibit systematic bias toward confusing specific classes. Despite the underlying class imbalance, where certain attraction levels are more represented than others, the model effectively handles this asymmetry, demonstrating robustness and reliability in distinguishing between different levels of attraction.

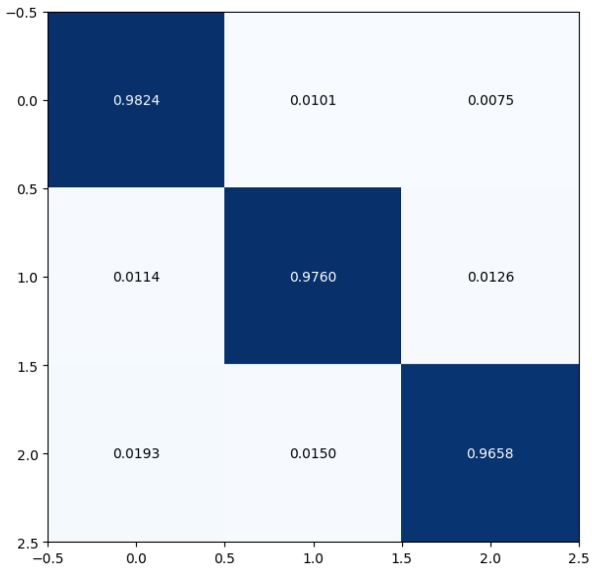

Figure 6: Normalized confusion matrix for attraction classification.

4.3. Task 3: Identifying the Associated Magical Town

Four classifier versions were developed and evaluated for 40 Magical Towns, considering class imbalance.

4.3.1. Classifier Performance

Table 3 summarizes the performance on the test set (41611 instances). The initial classifiers, based solely on the `Magical Towns` dataset, performed poorly. TF-IDF achieved an accuracy of 0.39 and an F1-score of 0.46, limited by the lack of neighbors. BETO with UMAP was ineffective (accuracy 0.02, F1-score 0.29), unable to generalize with one instance per class. RAKE with TF-IDF performed even worse (accuracy 0.081, F1-score 0.087), losing crucial context.

Method	**Accuracy**	**Weighted F1-score**
TF-IDF + KNN	0.39	0.46
BETO + UMAP + KNN	0.02	0.29
RAKE + TF-IDF + KNN	0.081	0.087
TF-IDF + KNN	0.58	0.57

Table 3: Classification results on the test set.

Combining `Magical Towns` with `Reviews` increased instances per class, improving the classifier. TF-IDF with `cosine` and $k = 15$ achieved an accuracy of 0.5863 and an F1-score of 0.5762 on a test set of 41,611 instances. Fig. 7 shows the normalized confusion matrix, reflecting the improvement of the final classifier.

The final model showed notable improvements, with majority classes like Tulum and Isla Mujeres performing well, while minority classes like Tapalpa and Tepotzotlán had lower performance due to class imbalance, as shown in Table 4. Some towns with low support, such as Chiapa de Corzo, Xilitla, Real de Catorce, and Cuatro Ciénegas, achieved high F1-scores, likely due to distinctive review terms that TF-IDF effectively weighted, ensuring accurate classification despite limited data. High precision for these towns indicates reliable predictions, though lower recall suggests some missed instances. In contrast, towns with generic review content, like Tapalpa and Tepotzotlán, faced classification challenges. TF-IDF preprocessing, by removing stop-words and accents, likely enhanced distinctive terms for some towns while reducing context for others with less unique vocabulary.

5. Conclusions and Future Work

Regarding the polarity analysis task, the model achieved robust performance, with an overall accuracy of 76.84% and a macro F1-score of 0.6155. The results highlight strong performance on the majority class (polarity 4), which constitutes the bulk of the corpus, with high precision, recall, and F1-score values, confirming that the model effectively learned to identify frequently occurring positive opinions. However, minority classes, particularly the negative ones (0 and 1), exhibited considerably lower performance, revealing challenges in classifying underrepresented texts and a tendency to confuse adjacent classes with similar polarity (e.g., classes 2 and 3). This behavior clearly reflects the impact of the pronounced class imbalance, biasing the model toward the dominant category and limiting its capacity to discriminate subtle sentiment nuances.

Concerning the attraction identification task, the fine-tuned BETO model demonstrated excellent performance. By preserving the original class distribution during dataset splitting and employing an appropriate training setup, the model achieved high precision, recall, and F1-scores across all classes, with an overall accuracy of 97.61%. The confusion matrix further confirms the model's robustness, showing strong accuracy in classifying each attraction level and minimal, well-distributed misclassifications. These results highlight BETO's effectiveness and reliability in accurately distinguishing between different attraction categories despite class imbalance.

With respect to the Magical Town classification task, a KNN classifier was developed. The initial idea of using only Wikipedia descriptions resulted in poor performance due to the scarcity of data. The incorporation of preprocessed reviews and TF-IDF with the cosine metric (k = 15) improved performance (accuracy 0.58, F1-score 0.57), but class imbalance limited the model's generalization capability. The imbalance (Tulum with 21.79% vs. Tapalpa with 0.35%) favored the majority classes. Preprocessing may have removed relevant information, and k = 15 suggests that the classifier needed more neighbors to

Magical Town	Precision	Recall	F1-score	Support
Ajijic	0.55	0.43	0.48	751
Atlixco	0.45	0.40	0.43	289
Bacalar	0.39	0.62	0.48	2165
Bernal	0.52	0.37	0.43	251
Chiapa de Corzo	0.77	0.58	0.66	192
Cholula	0.55	0.46	0.50	558
Coatepec	0.68	0.43	0.53	163
Creel	0.55	0.41	0.47	357
Cuatro Ciénegas	0.87	0.41	0.56	157
Cuetzalan	0.65	0.23	0.34	199
Dolores Hidalgo	0.68	0.31	0.42	182
Huasca de Ocampo	0.72	0.46	0.56	302
Isla Mujeres	0.58	0.79	0.67	5965
Ixtapan de la Sal	0.61	0.41	0.49	339
Izamal	0.71	0.43	0.53	408
Loreto	0.71	0.45	0.55	1105
Malinalco	0.55	0.35	0.43	286
Mazunte	0.67	0.37	0.47	293
Metepec	0.29	0.48	0.36	707
Orizaba	0.65	0.44	0.53	504
Palenque	0.53	0.45	0.49	1903
Parras	0.68	0.37	0.48	191
Pátzcuaro	0.69	0.44	0.53	891
Real de Catorce	0.82	0.46	0.59	152
San Cristóbal de las Casas	0.46	0.62	0.52	2612
Sayulita	0.61	0.39	0.48	1467
Tapalpa	0.92	0.24	0.38	145
Taxco	0.68	0.49	0.57	840
Teotihuacán	0.74	0.61	0.67	1162
Tepotzotlán	0.66	0.16	0.26	203
Tepoztlán	0.54	0.39	0.45	689
Tequila	0.67	0.50	0.57	530
Tequisquiapan	0.39	0.30	0.34	725
Tlaquepaque	0.73	0.32	0.45	808
Todos Santos	0.81	0.33	0.47	920
Tulum	0.68	0.83	0.74	9069
Valladolid	0.58	0.46	0.51	2327
Valle de Bravo	0.57	0.37	0.45	1192
Xilitla	0.94	0.51	0.66	291
Zacatlán	0.84	0.34	0.48	321

Table 4: Classification report for TF-IDF with augmented data (`cosine`, $k = 15$).

compensate for the imbalance.

It is worth noting that all our models outperformed the baseline established by the challenge organizers. This outcome highlights the effectiveness of the proposed methodologies across the three tasks addressed. The improvements over the baseline underscore the relevance of our approaches and their suitability for Spanish-language text analysis.

For the tasks with the greatest class imbalance (Polarity and Identifying the Associated Magical Town),

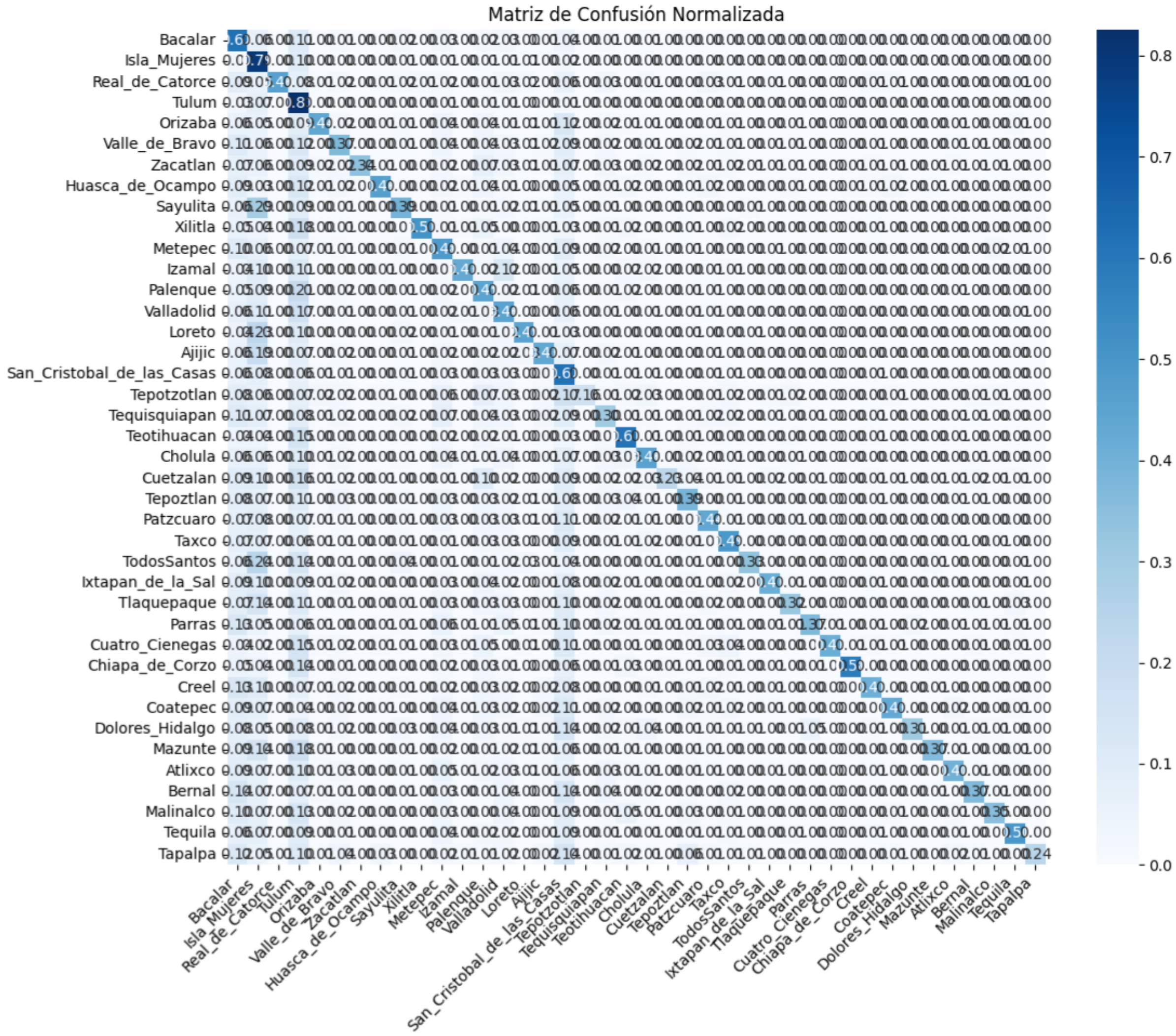

Figure 7: Normalized confusion matrix for TF-IDF with augmented data (cosine, $k = 15$).

the models proved effective in recognizing the predominant classes. However, the results highlight the need for future improvements, such as class balancing techniques. In particular, the generation of synthetic text using large language models like GPT is proposed to perform data augmentation with synthetic examples, aiming to enhance the representation of minority classes.

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] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (2014) 1093–1113.

[2] M. Haselmayer, M. Jenny, Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding, Quality & quantity 51 (2017) 2623–2646.

[3] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, H. Mora, Managing marketing decision-

making with sentiment analysis: An evaluation of the main product features using text data mining, Sustainability 11 (2019) 4235.
[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 (2023) 639–661. URL: http://dx.doi.org/10.1108/IJTC-09-2022-0223. doi:10.1108/ijtc-09-2022-0223.
[5] A. Giachanou, F. Crestani, Like it or not: A survey of twitter sentiment analysis methods, ACM Computing Surveys (CSUR) 49 (2016) 1–41.
[6] R. Guerrero-Rodríguez, M. A. Álvarez-Carmona, R. Aranda, et al., Big data analytics of online news to explore destination image using a comprehensive deep-learning approach: a case from mexico, Information Technology & Tourism 26 (2024) 147–182. URL: https://doi.org/10.1007/s40558-023-00278-5. doi:10.1007/s40558-023-00278-5.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] M. H. Abd El-Jawad, R. Hodhod, Y. M. Omar, Sentiment analysis of social media networks using machine learning, in: 2018 14th international computer engineering conference (ICENCO), IEEE, 2018, pp. 174–176.
[14] S. T. Kokab, S. Asghar, S. Naz, Transformer-based deep learning models for the sentiment analysis of social media data, Array 14 (2022) 100157.
[15] S. Mann, J. Arora, M. Bhatia, R. Sharma, R. Taragi, Twitter sentiment analysis using enhanced bert, in: Intelligent Systems and Applications: Select Proceedings of ICISA 2022, Springer, 2023, pp. 263–271.
[16] M. Singh, A. K. Jakhar, S. Pandey, Sentiment analysis on the impact of coronavirus in social life using the bert model, Social Network Analysis and Mining 11 (2021) 33.
[17] M. Graff, D. Moctezuma, E. S. Téllez, Bag-of-word approach is not dead: A performance analysis on a myriad of text classification challenges, Natural Language Processing Journal 11 (2025) 100154. URL: https://www.sciencedirect.com/science/article/pii/S2949719125000305. doi:https://doi.org/10.1016/j.nlp.2025.100154.
[18] R. Guerrero-Rodriguez, M. A. 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. URL: https://doi.org/10.1080/13683500.2021.2007227. doi:10.1080/13683500.2021.2007227. arXiv:https://doi.org/10.1080/13683500.2021.2007227.

[19] 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.

[20] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. URL: https://www.sciencedirect.com/science/article/pii/0306457388900210. doi:https://doi.org/10.1016/0306-4573(88)90021-0.

[21] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. URL: https://arxiv.org/abs/2308.02976. arXiv:2308.02976.

[22] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL: https://arxiv.org/abs/1802.03426. arXiv:1802.03426.

[23] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic Keyword Extraction from Individual Documents, 2010, pp. 1 – 20. doi:10.1002/9780470689646.ch1.