

Multi-Task Text Classification of Tourist Reviews Using Doc2Vec

Lorena Vazquez¹, Jean Arreola^{2,*}

¹National Polytechnic Institute, Mexico

²Center for Research in Mathematics, Mexico

Abstract

Understanding tourist experiences through online reviews offers valuable insights for enhancing destination management and marketing, particularly in culturally rich regions such as Mexico's Pueblos Mágicos. In this work, we present a multi-task natural language processing (NLP) approach to automatically analyze user-generated reviews by predicting three key attributes: sentiment polarity (on a 1–5 scale), review category (restaurant, hotel, or attractive), and the corresponding tourist destination. Our model leverages Doc2Vec, an unsupervised algorithm that learns fixed-length vector representations of variable-length texts, allowing us to capture semantic and contextual information from Spanish-language reviews. These document embeddings are then used as input features for a set of supervised classifiers tailored to each task. The proposed pipeline was developed and evaluated in the context of the Rest-Mex challenge, where multilingual, informal, and domain-specific language posed unique modeling challenges. Results demonstrate that Doc2Vec embeddings provide a strong foundation for robust document-level classification across multiple tasks, enabling scalable analysis of subjective tourist feedback. This research contributes a replicable framework for semantic understanding of tourism narratives in regional and linguistically diverse settings.

Keywords

Embedding, Mexican tourism, Magical Town, Doc2Vec

1. Introduction

Tourism today is shaped not just by brochures and guidebooks, but by the voices of travelers who share their experiences online. Reviews posted by tourists offer valuable insights into what they loved, what could be improved, and how they felt during their visit [1, 2, 3]. In Mexico, the “Pueblos Mágicos” (Magical Towns) program highlights destinations known for their cultural richness, natural beauty, and historical significance [4]. Understanding how visitors talk about these places can help improve the tourist experience and support local businesses and communities.

As part of the REST-MEX [5, 6, 7, 8] [9] event, it was developed a system composed of two machine learning models and an heuristic, designed to analyze tourist reviews and extract key information automatically. The system focuses on three main tasks: (1) identifying the sentiment or “polarity” of a review on a scale from 1 (very negative) to 5 (very positive), (2) recognizing the type of review—whether it’s about a restaurant, a hotel, or a local attraction—and (3) determining which Pueblo Mágico the review refers to.

To achieve this, natural language processing (NLP) techniques were employed to train the system to understand and interpret text written in Spanish, which is often informal and rich in regional expressions. By combining linguistic analysis with machine learning, the system is able to process thousands of reviews quickly and consistently, providing a powerful tool for decision-makers in the tourism sector.

This paper explains how the system was built, the challenges with the data, and how the results can help guide smarter tourism strategies that reflect the real voices of travelers.

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

✉ lvazquez@alumno.ipn.mx (L. Vazquez); jean.arreola@cimat.mx (J. Arreola)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Dataset

The dataset used in this work was provided by the contest organizers. It consists of 208,051 reviews, each consisting of:

- Title: The title that the tourist assigned to their opinion
- Review: The opinion issued by the tourist.
- Polarity: The label representing the sentiment polarity of the opinion (1, 2, 3, 4, 5).
- Town: The town where the review is focused.
- Region: The region (state in Mexico) where the town is located.
- Type: The type of place the review refers to (Hotel, Restaurant, Attractive).

Polarity	Instances
1	5,441
2	5,496
3	15,519
4	45,034
5	136,561

Table 1

Total Instances of Polarity

Class	Instances
Hotel	51,410
Restaurant	86,720
Attractive	69,921

Table 2

Total Instances of Type

Town	Instances
Tulum	45,345
Isla Mujeres	29,826
San Cristóbal de las Casas	13,060
Valladolid	11,637
Bacalar	10,822
Palenque	9,512
Sayulita	7,337
Valle de Bravo	5,959
Teotihuacan	5,810
Other	68,743

Table 3

Total 10 selected Towns

3. Modelling Approach

3.1. Data preprocessing

A critical step in analyzing the reviews involved understanding their structure and identifying potential factors that could impact the embeddings, such as multilingual content, spelling errors, and inconsistent use of uppercase and lowercase letters.

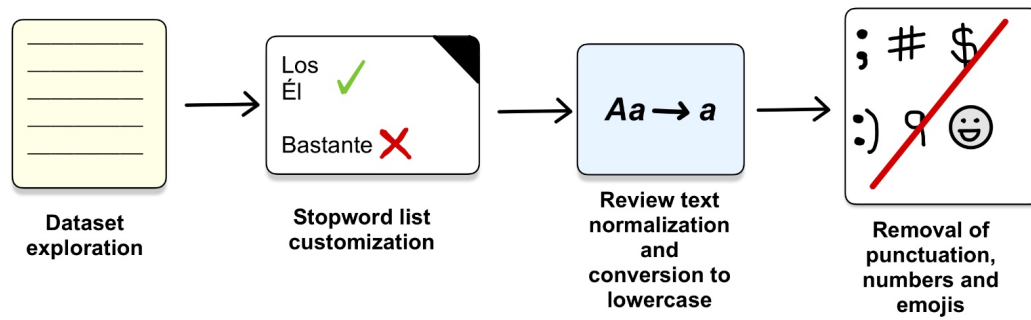


Figure 1: Preprocessing of the data.

To mitigate these issues, a series of data cleaning and preprocessing steps were applied (similar to [10]):

1. Selection of an appropriate file reading and text encoding method.
2. Customization of the stopwords list by modifying spaCy's [11] Spanish stopwords set, excluding adverbs, adjectives, verbs, and auxiliary verbs.
3. Standardization of spelling and conversion of all text to lowercase.
4. Removal of punctuation, numbers, and emojis, along with normalization of whitespace.

3.2. Doc2Vec

Doc2Vec [12] is an extension of the Word2Vec framework, designed to learn fixed-length vector representations for variable-length pieces of text such as sentences, paragraphs, or entire documents. It was introduced by Le and Mikolov in 2014 under the name Paragraph Vector.

Whereas Word2Vec learns word embeddings based on local context windows (using models such as CBOW or Skip-gram), Doc2Vec adds an additional mechanism to capture the semantics of larger text units by incorporating a document identifier or tag during training. The resulting vector embeddings capture not only word-level semantics but also higher-level structure and meaning across longer sequences.

Doc2Vec comes in two primary variants:

Distributed Memory Model of Paragraph Vectors (PV-DM): In this model, the document is mapped to a unique vector (document ID), which is concatenated or averaged with word vectors from the context window to predict the next word. This setup is analogous to the CBOW model in Word2Vec.

- It captures semantics and word order.
- The document vector acts as a sort of memory that retains what is missing from the current context.

Distributed Bag of Words version of Paragraph Vector (PV-DBOW): This model ignores the context words in the input and uses only the document vector to predict words sampled from the document.

- This approach is more akin to the Skip-gram model.
- It tends to capture topic-level information better.

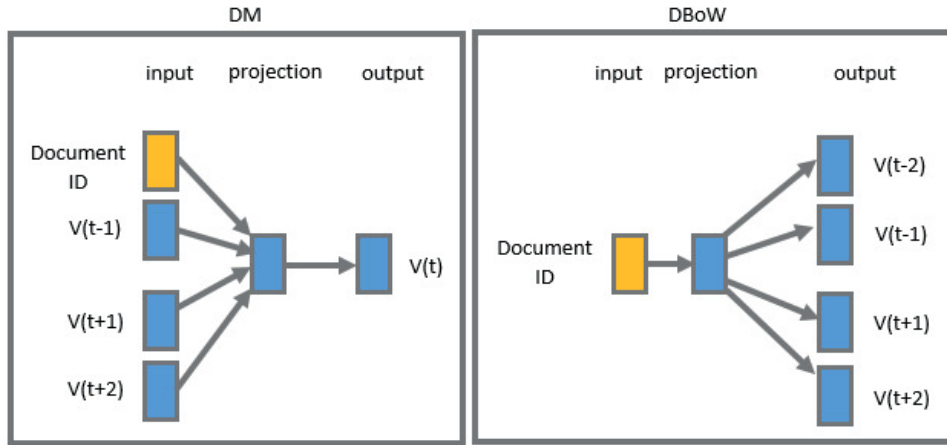


Figure 2: Doc2Vec has two main variants: Distributed Memory (DM) and Distributed Bag of Words (DBoW). In the DM model, during training, the document ID and the surrounding context words are used to predict a target word. In contrast, the DBoW model relies solely on the document ID to predict a set of words that appear within the document. [Reproduced from ResearchGate], via ResearchGate. (<https://www.researchgate.net/profile/Metin-Bilgin/publication/320829283/figure/fig3/AS:759096500297730@1557994117390/Doc2Vec-DM-and-DBoW.ppm>).

- Faster and simpler than PV-DM, but potentially less precise for fine-grained semantics.

In practice, both models are often trained together to leverage their complementary strengths.

In Doc2Vec, tags are essential because they uniquely identify each document in the training set. They serve two primary purposes:

- During training, each tag is associated with a trainable vector that is updated alongside word vectors. This tag vector becomes the document embedding.
- During inference, a new document (unseen during training) can be assigned a new tag, and its vector can be inferred by fixing the word vectors and optimizing only the new document vector.

Tags are not limited to single IDs—they can also include multiple labels or metadata (e.g., genre, author, source) if richer supervision is needed [13]. In the case of the polarity and type model, tags were obtained for each of the polarity and types categories.

Finally, it is important to highlight the advantages of the Doc2Vec model over its predecessor approaches, as it has demonstrated strong performance in capturing the semantic structure of texts.

3.3. Polarity Model

In this study, document embeddings for each review were generated using the Doc2Vec model. Additionally, the Type, Town, Region, and Polarity attributes were incorporated as tags.

For the classification task, an XGBoost model was trained using Polarity as the target variable. Hyperparameter optimization was carried out with the Optuna [14] library, with the objective of addressing class imbalance—an essential factor in enhancing model performance metrics.

3.4. Type of Review Model

For this model, document embeddings were obtained for each review, and for each "tag" of the type of review ("Restaurant", "Hotel", "Attractive"). Similarity was obtained between the review embedding,

and each of the tag embedding. The three similarities were concatenated to the review embedding.

A xgboost model was trained using the class weight of each of the categories [15].

3.5. Place Recognition Model

For the place recognition Model, a classification model was discarded, because of the number of categories.

An heuristic approach was used, following the next steps:

- A curated list of tourist destinations was compiled based on selections from [16], as well as recommendations generated by a GPT model specifically trained to retrieve information about Mexico’s Magical Towns. This model provided not only key attractions of each location but also commonly recommended hotels and restaurants. Additionally, the list included the colloquial or promotional nicknames associated with each destination. Consequently, when a particular location was referenced in a review, it was matched with the corresponding Magical Town based on these associations.
- Average embeddings for each type of review were obtained for every town.
- Similarity between the review embedding and all the place-type average embeddings was obtained, accordingly to the type of the review. The town with the highest similarity was matched.

4. Results and Discussion

The results of the systems and the baseline of the contest are presented in the tables below.

System	Macro F1	Accuracy	Avg Precision	Avg Recall	MAE
Polarity Model	0.19	36.91	0.2	0.2	1.044
Baseline Model	0.158	65.53	0.2	0.131	0.55
Top Ranked Model	0.987	78.52	0.62	0.66	0.22

Table 4
Contest overall results for Polarity Task

System	Macro F1	Accuracy	Avg Precision	Avg Recall
Type Model	0.358	37.53	0.358	0.36
Baseline Model	0.196	41.86	0.333	0.139
Top Ranked model	0.98	98.82	0.98	0.98

Table 5
Contest overall results for Type Task

System	Macro F1	Accuracy	Avg Precision	Avg Recall
Town Model	0.309	41.63	0.372	0.312
Baseline model	0.008	21.79	0.025	0.005
Top Ranked model	0.69	77.03	0.65	0.73

Table 6
Contest overall results for Town Task

The results obtained across the three tasks of the Rest-Mex task (polarity detection, review type classification, and place recognition) highlight both the strengths and limitations of our proposed

models in comparison to the Top Ranked Model.

In the polarity detection task, our model slightly surpassed the baseline in terms of Macro F1 and average recall. However, its mean absolute error (MAE) was notably high, and performance remained distant from that of the Top Ranked Model, which achieved near-perfect metrics. This discrepancy suggests that our approach may be overly sensitive to class imbalance or error propagation in sentiment boundaries. The high MAE also indicates that misclassifications were often far from the true label in ordinal space, suggesting the need to explore ordinal regression or class-weighted losses to better align training objectives with evaluation metrics.

For the review-type classification task, our model showed solid improvement over the baseline, especially in Macro F1 and recall. Yet, its performance still fell short of the Top Ranked Model, which achieved almost flawless results. This sharp contrast implies that our current feature representations and modeling strategy may be insufficient to capture the nuanced semantic distinctions between review types. The superior performance of the Top Ranked Model hints at the potential benefits of using domain-adapted language models, attention mechanisms, or even hierarchical classification architectures.

The place recognition task (Town Model) posed the greatest challenge, as reflected by the modest accuracy and F1 score achieved. Although our model outperformed the baseline, it was substantially outperformed by the top-ranked model. Place names in user-generated content are often noisy, misspelled, or contextually ambiguous, which may explain our model’s limited ability to disambiguate them effectively. Potential improvements include integrating spelling normalization, geographical knowledge bases, and context-aware entity linking methods that can better resolve ambiguous mentions.

Across all tasks, the Top Ranked Model established a clear performance ceiling, demonstrating the effectiveness of advanced modeling strategies. Their results suggest that a more holistic approach—including pre-trained transformer models, feature-rich inputs, and possibly ensemble techniques—is critical for achieving state-of-the-art performance in this domain.

5. Conclusions

Our models for the Rest-Mex challenge represent a significant step forward from baseline systems, particularly in tasks involving sentiment and type classification. However, the performance gap relative to the Top Ranked Model highlights key areas where our current pipeline could be improved.

Moving forward, we identify the following priorities.

- Mitigating class imbalance [15], especially in polarity prediction, through loss reweighting, data augmentation, or sampling strategies.
- Adopt transformer-based embeddings fine-tuned on tourism-specific corpora to improve semantic understanding.
- Incorporating domain knowledge (e.g., gazetteers, spatial hierarchies) to enrich context in place recognition tasks.
- Leveraging interpretability tools to analyze model errors and iteratively refine feature sets.

Exploring multitask and ensemble learning frameworks to capture the interdependence between tasks and diversify model reasoning.

In summary, while our models demonstrate encouraging progress, achieving top-tier performance in review understanding for tourism contexts will require a richer integration of linguistic, geographic, and contextual signals, supported by more sophisticated and adaptive learning architectures.

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-Rodríguez, 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>.
- [2] 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.
- [3] 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>.
- [4] J. L. García, M. d. C. Hernández, El estudio de los pueblos mágicos: Una revisión a casi 20 años de la implementación del programa, *Dimensiones Turísticas* 5 (2021) 9–38. URL: <https://dimensionesturisticas.amiturismo.org/wp-content/uploads/2021/04/DT-V5N8-Art-1-El-estudio-de-los-pueblos-magicos-9-38.pdf>.
- [5] 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>.
- [6] 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).
- [7] 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).
- [8] 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.
- [9] 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.
- [10] J. Arreola, L. Garcia, J. Ramos-Zavaleta, A. Rodriguez, An embeddings based recommendation system for mexican tourism. submission to the rest-mex shared task at iberlef 2021 (2021).

- [11] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, Zenodo (2020). URL: <https://doi.org/10.5281/zenodo.1212303>. doi:10.5281/zenodo.1212303.
- [12] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning (ICML), PMLR, 2014, pp. 1188–1196.
- [13] S. Chen, A. Soni, A. Pappu, Y. Mehdad, Doctag2vec: An embedding based multi-label learning approach for document tagging, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 111–120. URL: <https://aclanthology.org/W17-2614/>. doi:10.18653/v1/W17-2614.
- [14] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), ACM, 2019, pp. 2623–2631. URL: <https://arxiv.org/abs/1907.10902>. doi:10.1145/3292500.3330701.
- [15] A. Allawala, A. Ramteke, P. Wadhwa, Performance impact of minority class reweighting on xgboost-based anomaly detection, International Journal of Machine Learning and Computing 12 (2022) 143–148. URL: <http://www.ijmlc.org/vol12/1093-IJMLC-3012.pdf>. doi:10.18178/ijmlc.2022.12.4.1093.
- [16] Secretaría de Turismo de México, Pueblos mágicos, <https://www.sectur.gob.mx/gobmx/pueblos-magicos/>, 2025. Accessed: 2025-04-01.