A test of Mutual Information Features in Multi-Task Classification Spanish Tourist Reviews

Alejandra Romero-Canton[1,*], Jose Ramon Aranda-Romero[1]

[1]Secretaría de Educación del Gobierno del Estado de Yucatán, Mérida, México

Abstract

We propose a Mutual Information (MI)-based framework for balanced classification of Spanish-language tourist reviews using the Rest-Mex 2025 dataset. Our system predicts sentiment polarity, business type (hotel, restaurant, attraction), and geographic location (state and municipality) from over 200,000 annotated entries. We address class imbalance through redundancy pruning and synthetic data augmentation while weighting feature tokens using normalized MI scores. These scores are computed across class labels to capture the discriminative power of each term. Combined with FastText classifiers and rich preprocessing pipelines, our MI-driven approach improves fairness, interpretability, and accuracy in multi-label tourism classification tasks. Results show strong performance in business type classification (F1 = 0.9687) and improved balance across minority classes. This work highlights the potential of combining statistical information-theoretic measures with modern NLP pipelines for real-world tourism sentiment analysis.

Keywords

Mutual Information, Sentiment Analysis, Rest-Mex

1. Introduction

Travel & Tourism Competitiveness Report (TTCR) [1], published by the World Economic Forum, the Travel & Tourism (T&T) industry was highlighted as undergoing significant expansion. According to the World Tourism Organization (UNWTO), global international tourist arrivals had reached 1.4 billion in 2018—an achievement that surpassed previous forecasts by two years. Nonetheless, the TTCR also cautioned that unchecked growth and competitiveness could jeopardize the very resources that sustain the sector.

Two years later, the outlook of the T&T industry shifted dramatically. The COVID-19 pandemic dealt a severe blow to travel demand, causing widespread disruption through lockdowns, travel bans, and the collapse of international mobility. These effects were not limited to businesses but extended to economies reliant on tourism. Although signs of recovery are now evident, they vary notably across regions and markets. The path forward is further complicated by global events such as the war in Ukraine [2].

These disruptions have likely led to lasting transformations in both the industry and traveler behavior. Tourists are now more attentive to health and sanitation at destinations, and remain wary of potential COVID variants, regulatory changes, and travel interruptions. The pause in global travel has also encouraged reflection on the environmental impact of tourism. In response, both governments and tourism enterprises have begun reevaluating their strategies, reallocating investments, and implementing measures to better manage risk and evolving consumer expectations.

Beyond the pandemic, the tourism sector has undergone a technological shift over the past decade. Innovations in digitization, information and communication technologies, machine learning, robotics, and artificial intelligence (AI) have reshaped the way travelers interact with destinations [3, 4, 5, 6, 7, 8]. Today, most international travelers rely on digital platforms to plan their trips, with online information playing a significant role in their decision-making process [9, 10, 11].

IberLEF 2025, September 2025, Zaragoza, Spain

[*]Corresponding author.

ale_romero85@hotmail.com (A. Romero-Canton)

Sentiment analysis and classification of user-generated content in tourism offer significant insights into customer satisfaction, regional trends, and service quality. However, data imbalance and noisy labels present major challenges. While recent NLP pipelines leverage embeddings and deep learning, they often underperform for low-resource classes. This paper presents a hybrid approach that combines Mutual Information (MI)-based feature weighting with class balancing to enhance classification robustness on the Rest-Mex 2025 corpus.

2. Dataset Description

Unlike the past editions [12, 13, 14], the Rest-Mex 2025 corpus [15, 16], containing 208,051 reviews labeled with:

- **Sentiment Polarity**: Ordinal scale from 1 (very negative) to 5 (very positive).
- **Business Type**: Hotel, Restaurant, Attraction.
- **Geographic Location**: State and municipality in Mexico.

The dataset is multilingual, domain-specific, and imbalanced, with towns like Tulum overrepresented.

3. Methodology

3.1. Text Preprocessing

Each review is normalized using a custom class, including:

- Lowercasing, whitespace trimming.
- Stopword removal (extended Spanish list).
- Digit normalization via semantic character replacement.
- Tokenization (NLTK), Lemmatization (spaCy).

3.2. Mutual Information Feature Extraction

The main contribution of this stage, compared to previous work such as [17], lies in how we extract informative features for each classification subtask (i.e., sentiment polarity, business type, and geographic location). Inspired by earlier applications of Mutual Information (MI) in NLP [18, 19], we leverage MI to quantify the association between words and class labels across the Rest-Mex 2025 dataset.

Mutual Information measures the shared information between two variables, X and Y, and is defined as:

$$MI(X, Y) = P(X, Y) \log \left(\frac{P(X, Y)}{P(X)P(Y)} \right) \tag{1}$$

where $P(X, Y)$ denotes the joint probability of observing word X in class Y. When X and Y are independent, $MI(X, Y) \approx 0$, implying that X carries no useful signal for predicting Y. In contrast, high MI values indicate strong association, making such words discriminative for their corresponding classes [20].

In our application, each token $b \in B$ (where B is the vocabulary of the corpus) is evaluated against each class $c \in C$ (e.g., sentiment levels or location labels):

- If word b occurs uniformly across all classes, $MI(b, c) \approx 0$, offering little to no discriminative power.
- If b is mostly exclusive to class c, then $MI(b, c) > 0$, suggesting strong relevance.
- If b appears frequently in other classes but rarely in c, then $MI(b, c) < 0$, indicating negative association.

To enhance the representational power of each class, we expand the high-MI words by including up to five synonyms retrieved via WordNet, assigning each synonym the same MI score as the original word. This results in a trained feature set Ω_c for class c, where each element is a tuple $\omega_{i,c} = (\omega^{w}_{i,c}, \omega^{MI}_{i,c})$, representing a word and its normalized MI score. This enriched feature space captures both statistical relevance and semantic diversity, providing a more robust foundation for classification [21].

3.3. Balancing Strategy

For overrepresented classes: remove redundant reviews via hybrid Jaccard/Fuzzy similarity. For underrepresented classes: generate synthetic reviews using MI-rich vocabulary templates. All classes are equalized to the global mean number of instances $\bar{n}$.

3.4. Model Training and Prediction

We train three FastText classifiers on balanced, MI-weighted inputs:

- Polarity model (5-class ordinal).
- Business type model (3-class nominal).
- Location model (40-class multiclass).

Prediction follows softmax scoring, with highest probability label chosen per model.

4. Results and Evaluation

Table 1 shows macro-averaged metrics on a test set (20% held-out):

Task	Accuracy	F1	Precision	Recall
Business Type	0.6696	0.6687	0.6690	0.6684
Location	0.4657	0.4397	0.4172	0.4771
Polarity	0.5905	0.4403	0.4824	0.4402

Table 1
Overall classification performance.

MI-based features can not improved minority class performance either reduced confusion in adjacent polarities.

5. Conclusions

In this work, we proposed a Mutual Information-driven approach for balanced text classification in the context of the Rest-Mex 2025 challenge. By integrating MI-based feature selection with strategic data balancing—through redundancy reduction and synthetic augmentation—we aimed to mitigate the limitations caused by class imbalance and noisy inputs. The MI scores helped us identify discriminative terms for each class, improving interpretability and enabling targeted vocabulary expansion via synonym enrichment.

Compared to previous MI-based models, our methodology benefits from a more refined preprocessing pipeline and multi-label FastText classifiers tailored for sentiment, business type, and geographic prediction. The results show low performance on the business classification task and acceptable results in more complex subtasks like location and sentiment polarity. In the context of this task, MI is not enough for solving problems with many data, as it gets confused.

Nonetheless, challenges remain. Some tokens with high MI values were semantically irrelevant or derived from noisy user-generated content (e.g., *queretarcdm*, *metrocdmx*). Future work will focus on automatic filtering of such terms, integrating contextual embeddings, and refining semantic augmentation strategies to further improve robustness and fairness across all 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] L. U. Calderwood, M. Soshkin, The travel and tourism competitiveness report 2019, 2019.

[2] Travel & tourism development index 2021, rebuilding for a sustainable and resilient future, 2022.

[3] R. T. Qiu, J. Park, S. Li, H. Song, Social costs of tourism during the covid-19 pandemic, Annals of Tourism Research 84 (2020) 102994. URL: https://www.sciencedirect.com/science/article/pii/S0160738320301389. doi:https://doi.org/10.1016/j.annals.2020.102994.

[4] S. Gossling, D. Scott, C. M. Hall, Pandemics, tourism and global change: a rapid assessment of covid-19, Journal of Sustainable Tourism 29 (2021) 1–20. URL: https://doi.org/10.1080/09669582.2020.1758708. doi:10.1080/09669582.2020.1758708. arXiv:https://doi.org/10.1080/09669582.2020.1758708.

[5] J. Guerra-Montenegro, J. Sanchez-Medina, I. Lana, D. Sanchez-Rodriguez, I. Alonso-Gonzalez, J. Del Ser, Computational intelligence in the hospitality industry: A systematic literature review and a prospect of challenges, Applied Soft Computing 102 (2021) 107082. URL: https://www.sciencedirect.com/science/article/pii/S1568494621000053. doi:https://doi.org/10.1016/j.asoc.2021.107082.

[6] D. Buhalis, Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article, Tourism Review 75 (2020) 267–272. URL: https://doi.org/10.1108/TR-06-2019-0258. doi:10.1108/TR-06-2019-0258, publisher: Emerald Publishing Limited.

[7] A. Diaz-Pacheco, M. A. 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 0 (2022) 1–31. doi:10.1080/0952813X.2022.2153276.

[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] F. A. C. Calderón, M. V. V. Blanco, Impacto de internet en el sector turístico, Revista UNIANDES Episteme 4 (2017) 477–490.

[10] 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.

[11] 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.

[12] 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.

[13] 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: Recommen-

dation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[14] 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).

[15] 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.

[16] 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.

[17] A. Romero-Cantón, R. Aranda, AngelDiaz-Pacheco, J. P. Ramírez-Silva, Mexican epidemiological semaphore color prediction by means of mutual information features, in: CEUR Workshop Proceedings, Coruña, Spain, 2022.

[18] 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 (2021) 1–16. doi:https://doi.org/10.1080/13683500.2021.2007227.

[19] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, arXiv preprint arXiv:1808.06670 (2018).

[20] M. Ravanelli, Y. Bengio, Learning speaker representations with mutual information, arXiv preprint arXiv:1812.00271 (2018).

[21] M. Á. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, Journal of Intelligent & Fuzzy Systems 34 (2018) 2983–2990. doi:10.3233/JIFS-169483, publisher: IOS Press.