Rest-Mex 2025:
Sentiment Analysis and Magical Towns Detection Task

Alejandro Hernández-Baca[1,†], Miguel Ángel Rojas-Andrade[1,†],
Jessica Nohemí Figueroa-Ramírez[1,†] and José Roberto Prieto-Valdivia[*,†]

[1] *University of Guanajuato Department of Electronics Engineering, Carretera Salamanca – Valle de Santiago, km 3.5 + 1.8, Comunidad de Palo Blanco, 36787 Salamanca, Guanajuato, Mexico.*

Abstract

A predictive model for text classification was developed using Natural Language Processing (NLP) techniques, incorporating the pre-trained language model BERT Mini and machine learning classifiers. The model was applied to classify reviews from the TripAdvisor platform into three categories: sentiment polarity, city, and type of tourist attraction. Additionally, a genetic algorithm was used to select the most relevant features from the embeddings obtained from BERT Mini. The classification was performed using Self-Organizing Maps (SOM), and the results showed macro F1-scores of 0.50 for polarity, 0.28 for city, and 0.91 for type of attraction, highlighting the model's strong performance in attraction classification.

1. Introduction

In this work, we present the participation of the NLPnudos team in the REST-MEX 2025 competition [1, 2], unlike past editions [3, 4, 5]which aimed to address three tasks using tourism reviews: (1) detection of sentiment polarity, (2) categorization of the type of tourist destination, and (3) identification of the place to which the review belongs. This challenge focuses on the application of Natural Language Processing (NLP) techniques applied to the automated analysis of tourism texts in Spanish [6]. Tourism represents one of the most important economic sectors for Mexico, contributing 8.6% of the national Gross Domestic Product (GDP) [7]. Regarding international tourism, Mexico recorded a 10.5% increase in tourist arrivals during the January to May period between 2022 and 2023, generating a record economic income of 30.81 billion USD in foreign exchange [8]. In this context of high economic relevance and massive digitalization, Natural Language Processing (NLP) has become a key tool to extract knowledge from large volumes of user-generated opinions[9, 10, 11, 12]. One of the most established tasks in this field is sentiment analysis, whose goal is to automatically classify a review as positive, negative, or neutral. Models based on Bidirectional Gated Recurrent Units (BiGRU), along with attention mechanisms, have been shown to achieve accuracies above 90% in this task when applied to texts in the tourism domain [13].

Additionally, NLP has been implemented to classify the type of tourist destination, such as hotels, restaurants, parks, or cultural sites, through multi-class classification schemes. Recent works have proposed hybrid architectures that integrate multiple classifiers, even surpassing the performance of BERT-based models in specific categorization tasks [14]. In the current scenario, various implementations can be applied to different aspects of tourism, such as systems for cataloging dishes in restaurants, enabling a classification method to assess the quality of the dishes offered [15].

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

✉ a.hernandezbaca@ugto.mx (A. Hernández-Baca); ma.rojasandrade@ugto.mx (M. ∶ Rojas-Andrade); jn.figueroaramirez@ugto.mx (J. N. Figueroa-Ramírez); jr.prietovaldivia@ugto.mx (J. R. Prieto-Valdivia)
🌐 https://github.com/Alejandro-HB (A. Hernández-Baca)

published 2025-12-10

2. Related Work

In sentiment polarity analysis, models based on machine learning and deep learning have been proposed. For example, in [16] Li et al. employed a Bidirectional Recurrent Neural Network (BiRNN) that captured contextual dependencies to classify tourism reviews, achieving competitive performance. Other approaches, such as Aspect-Based Sentiment Analysis, enabled the identification of polarity associated with specific attributes mentioned in the text. Additionally, large-scale models like LLMs have recently been applied in domains such as finance, demonstrating their generalization capability for similar tasks (LLMs and NLP Models in Cryptocurrency) [17]. Regarding the classification of destination types, some methods have been developed to infer the category to which a review belongs (hotel, restaurant, park, etc.). In BERT-based Tourism Named Entity Recognition, the authors applied a BERT-based architecture to identify tourism-related entities in texts from social media, highlighting its ability to capture relevant names and categories in natural language. On the other hand, Tourism Profiling: A Semi-Automatic Classification Model of Points of Interest proposed a semi-automatic approach using an SVM classifier, demonstrating improved performance compared to traditional models in the categorization of points of interest [18].

These studies have shown encouraging results; however, many focus on general contexts or other languages, which underscores the need to explore approaches adapted to data in Spanish and specifically oriented towards tourism in Mexico, as proposed in this study.

3. Data Analysis

The dataset consists of a training collection with 208,051 instances containing data in Spanish, including the following columns: Title, Review, Polarity, Town, and Type; where the last three correspond to the target classes to be predicted. Initially, the data were examined to assess class balance. Due to a significant imbalance in the polarity classes—where the majority class exceeded half of the total samples—it was decided to downsample the data. Specifically, the number of samples for the dominant class was reduced to approximately match the size of the smallest class. Consequently, for polarity classification, around 6,000 samples per class were randomly selected. Similarly, for the city classification task, approximately 1,000 samples per class were retained. Finally, for the attraction type classification, about 60,000 samples were used, distributed across the three classes. Thus, three separate datasets were prepared, each with the objective of predicting polarity, city, and attraction type, respectively.

4. Methodology

Having partitioned the original dataset, a genetic algorithm was implemented for feature selection and evaluated using various classifiers. Random Forest showed the best performance among them; however, its F1-score did not exceed 0.5 in all cases. Therefore, an alternative approach was sought while still leveraging the feature selection, which successfully reduced the feature set from the original 384 features output by BERT Mini to approximately 230 features per dataset.

The chosen alternative was the Self-Organizing Map (SOM), considered a straightforward option for classification based on nearest neighbors. A SOM with a 100x100 grid was proposed, with the size experimentally chosen considering computational memory limitations. The training involved 150,000 epochs with a learning rate of 0.3. Although some regions of the map contained overlapping data points, the mode was applied within each cell to retain the most frequent class. During prediction, the zone of the map to which the input belonged was identified, followed by a neighborhood review to ensure the prediction's accuracy.

This entire process was applied to each target class, as illustrated graphically in Figure 1, where the results obtained with the SOM are presented later.

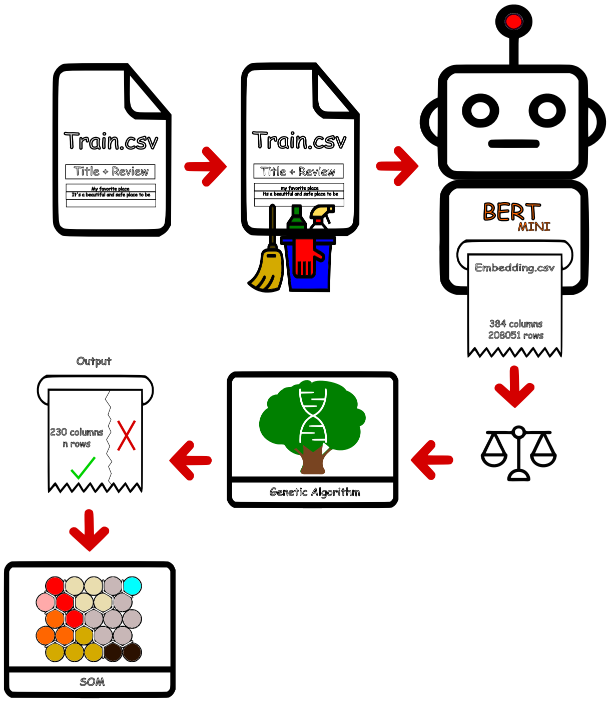

Figure 1: Methodology of the proposed model.

4.1. Processing Data

Data preprocessing consisted of a set of operations aimed to prepare, clean, and structure information to generate a suitable representation for subsequent processing by machine learning models. This step is crucial as it directly impacts the quality and performance of the model.

In the REST-MEX 2025 challenge, the inputs consisted of textual reviews of tourist destinations in Mexico. To enrich the semantic information available to the model, the review title was concatenated with its content, forming a single input sequence per instance. Subsequently, text cleaning was performed, which involved converting all text to lowercase, removing accents, punctuation marks, and numbers.

To reduce bias caused by class imbalance and improve the model's generalization ability, a sampling technique was applied that trimmed the dataset to a total of 208,051 reviews with titles. This adjustment allowed maintaining a balanced distribution among classes, preventing the model from predominantly favoring a single category during training.

4.2. BERT Mini model.

In the field of Natural Language Processing (NLP), Transformer-based models have revolutionized how tasks such as text classification, sentiment analysis, and Named Entity Recognition (NER) are approached. These models are characterized by their ability to capture long-range contextual dependencies in text sequences through self-attention mechanisms. However, their size and computational requirements vary considerably, making it necessary to select the appropriate model based on available resources and task complexity.

One of the most influential models in this category is BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin et al. in 2018. Unlike traditional sequential processing models, BERT is bidirectional, meaning it can understand the context of a word by considering both preceding and succeeding words in a sentence. This bidirectionality enables a deep contextual representation of language, which is key for semantic understanding tasks.

BERT Training: BERT was pretrained on a massive corpus of unlabeled text including the English Wikipedia (2.5 billion words) and the BookCorpus (800 million words).

4.3. Genetic Algorithm

A genetic algorithm was implemented to reduce the dimensionality of the features obtained from BERT Mini, which generates embeddings with 384 features that lead to high computational resource consumption.

The genetic algorithm aimed to select the most significant features for predicting the target class. For this purpose, the Random Forest algorithm was used as the classifier, and the F1-score was employed as the fitness metric due to the imbalanced nature of the dataset classes.

4.4. SOM

A 100x100 self-organizing map (SOM) was implemented for each class, yielding one prediction per class. In cases where class overlaps occurred—typically at the boundaries of the groups formed by the map—the statistical mode was applied among the overlapping classes within each cell to resolve conflicts, assigning the mode class to that cell. Subsequently, a second validation step was performed by examining the eight neighboring cells and applying the mode again, ensuring the most accurate prediction.

5. Experiments and Results

These results highlight the variability in the model's ability to handle different types of labels, likely influenced by data distribution and the inherent complexity of each task.

5.1. Feature Reduction (Genetic Algorithm)

The algorithm operates with populations of 20 individuals, each composed of a random selection of features in the initial generation, and continuously improves the fitness value over 50 generations. Other important hyperparameters are the crossover and mutation probabilities, set at 0.5 and 0.3, respectively.

It is important to mention that the Random Forest classifier can only handle one class for prediction at a time; therefore, the genetic algorithm must run separately for each class.

At the end of the genetic algorithm execution, a selection of 257 features was obtained for polarity, 321 for region, and 238 for attraction type.

5.2. Classification (SOM)

At the boundaries of the zones where the SOM separates classes, overlaps of multiple classes occur within the same cell. To resolve this and ensure each cell has a unique value, the mode is calculated among the overlapping classes, as illustrated in Figure 2. Subsequently, the mode is recalculated considering the eight neighboring cells surrounding the cell in question, as shown in Figure 3.

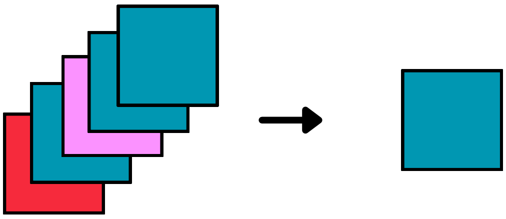

Figure 2: Mode applied in case of overlap.

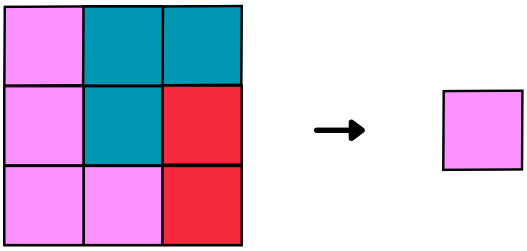

Figure 3: Mode applied with neighbors.

5.3. Metrics

The metric used to monitor the performance of the model was the F1 score. This metric is preferred over others, such as accuracy when dealing with imbalanced data because accuracy can give a misleading impression of good performance by favoring the majority class. In contrast, the F1-score combines precision and recall, providing a better reflection of how well the model correctly identifies the minority class, which is often the most important.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

5.4. Results

The results obtained for the three evaluated classes show differential behavior in the model's performance. For the Polarity class, the F1-score values remained around 0.49 across different tests, indicating moderate performance in identifying sentiment polarity. In the case of the Town class, the F1-score is considerably lower, ranging between 0.256 and 0.271, reflecting the greater difficulty of the model in correctly classifying the geographic regions or localities associated with the reviews. Finally, the Type class exhibits outstanding performance, with an F1-score exceeding 0.89, suggesting that the classification of the type of tourist attraction is much more accurate and stable compared to the other two categories.

In Figures 4, 5, and 6, we can observe how Fitness (F1-Score) improves over generations. This behavior reflects an optimization process in which the model achieves a more precise classification over time while simultaneously reducing the number of features required for effective classification. This phenomenon suggests that the memory usage in predictions could be optimized, as not all available data are essential for achieving optimal performance. Moreover, by reducing the number of features used, not only did the model's efficiency improve, but the computational cost associated with processing redundant information was also minimized. Consequently, this feature reduction strategy could be the key to developing lighter and more efficient models without compromising classification accuracy.

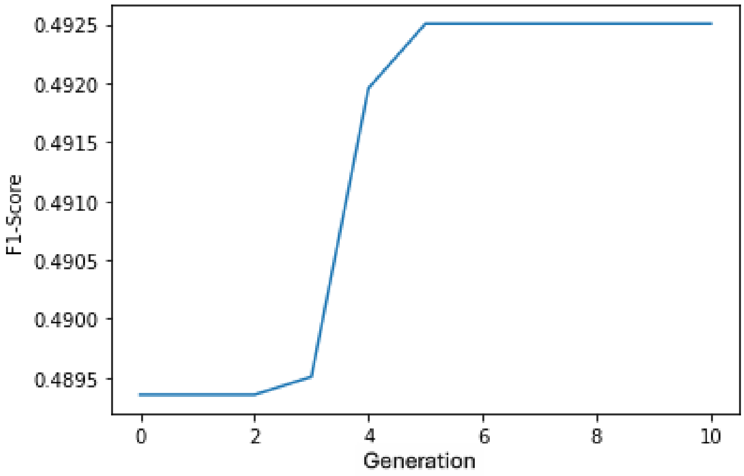

Figure 4: F1-Score plot over generations for the polarity label.

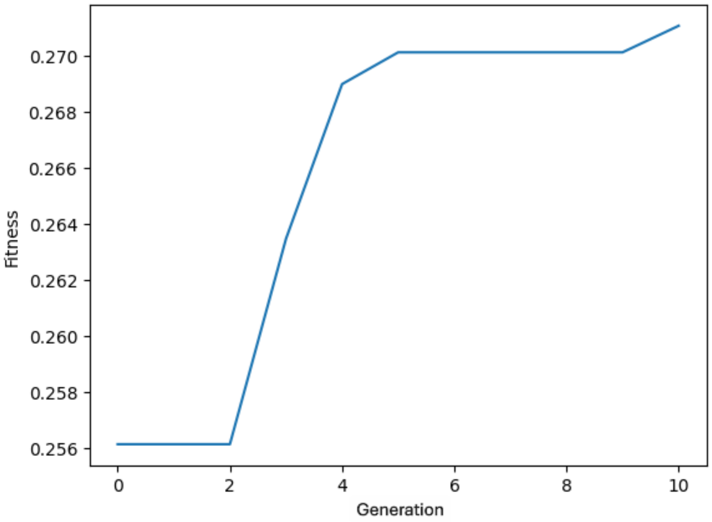

Figure 5: F1-Score plot over generations for the Town label.

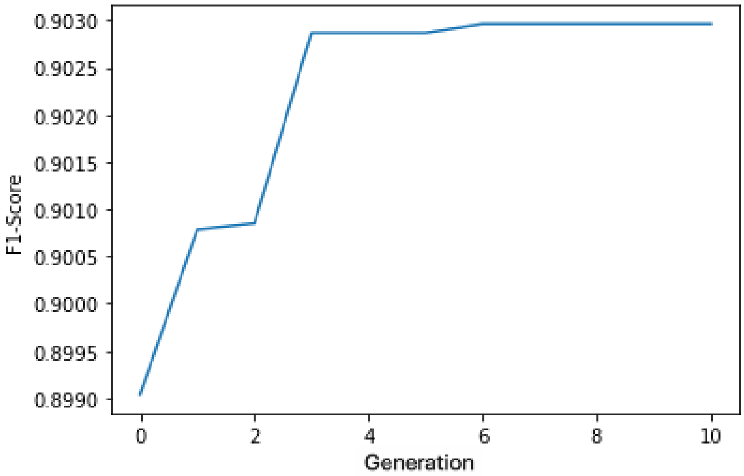

Figure 6: F1-Score plot over generations for the Type lable.

As observed in Tables 1 and 2, the F1-Score improved throughout the generations. However, by leveraging the reduced features, the SOM method achieved better performance in all classifications, which is why it was chosen to continue with this technique.

Table 1: Table of results F1-Score for generation.

	F1 - Score Generation 1	F1 - Score Generation 5	F1 - Score Generation 10
Polarity	0.489	0.492	0.492
Town	0.256	0.269	0.271
Type	0.899	0.902	0.902

Table 2: Table with results F1-Score SOM classification.

	F1 - Score SOM
Polarity	0.502
Town	0.283
Type	0.911

One persistent error observed in the classifier occurred when predicting the tourist's level of satisfaction with the destination. In several instances, the model assigned positive evaluations to entries that were, in fact, linked to clearly negative feedback. This misclassification may stem from a bias introduced during data preprocessing, particularly from balancing the dataset by trimming the surplus of positive examples. As a result, the model may have lost access to relevant signals for detecting dissatisfaction, leading it to overestimate actual satisfaction levels.

6. Conclusion

This work studied the classification of three classes —polarity, type, and town— using an NLP model based on Mini BERT, feature selection via a genetic algorithm, and classification through a Self-Organizing Map (SOM).

The results indicate that although the strategy of trimming classes to balance the dataset allowed for stable model training, this approach was insufficient to achieve high performance, especially for underrepresented classes such as Town, which yielded a notably low F1-score. Conversely, the Type class achieved significantly better performance, suggesting that data distribution and quantity greatly affect model effectiveness.

Therefore, it is concluded that simply reducing sample sizes to balance classes is not the most effective strategy for this type of problem. Instead, implementing data augmentation techniques to increase the diversity and number of samples in minority classes is recommended, which can enhance the model's generalization capability. Additionally, exploring the use of more powerful and robust models than BERT Mini, capable of capturing deeper and more discriminative language representations, is advisable to improve classification accuracy.

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] 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.

[2] 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.

[3] 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.

[4] 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).

[5] 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).

[6] 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, A. 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:10.1016/j.jksuci.2022.10.010.

[7] I. N. de Estadística y Geografía, CUENTA SATÉLITE DEL TURISMO DE MÉXICO (CSTM) 2023, 2024. URL: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2024/CSTM/CSTM2023.pdf?utm_source=chatgpt.com.

[8] Study on Accommodation Efficiency in "Pueblos Mágicos", Mexico: An Application of Data Envelopment Analysis (DEA) | IIETA, ???? URL: https://iieta.org/journals/mmc_d/paper/10.18280/mmc_d.451-405. doi:10.18280/mmc_d.451-405.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] T. Mou, H. Wang, Online comments of tourist attractions combining artificial intelligence text mining model and attention mechanism, Scientific Reports 15 (2025) 1121. URL: https://www.nature.com/articles/s41598-025-85139-3. doi:10.1038/s41598-025-85139-3, publisher: Nature Publishing Group.

[14] Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec | PLOS One, ???? URL: https://journals.plos.org/plosone/article?id=10.1371/journal.

pone.0305095.

[15] I. Castillo-Ortiz, M. A. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, Evaluating culinary skill transfer: A deep learning approach to comparing student and chef dishes using image analysis, International Journal of Gastronomy and Food Science 38 (2024) 101070. URL: https://www.sciencedirect.com/science/article/pii/S1878450X24002038. doi:10.1016/j.ijgfs.2024.101070.

[16] Q. Li, S. Li, J. Hu, S. Zhang, J. Hu, Tourism Review Sentiment Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism and Topic-Enriched Word Vectors, Sustainability 10 (2018) 3313. URL: https://www.mdpi.com/2071-1050/10/9/3313. doi:10.3390/su10093313, number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[17] N. Kumar, B. R. Hanji, Aspect-based sentiment score and star rating prediction for travel destination using Multinomial Logistic Regression with fuzzy domain ontology algorithm, Expert Systems with Applications 240 (2024) 122493. URL: https://www.sciencedirect.com/science/article/pii/S0957417423029950. doi:10.1016/j.eswa.2023.122493.

[18] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study, Big Data and Cognitive Computing 8 (2024) 63. URL: https://www.mdpi.com/2504-2289/8/6/63. doi:10.3390/bdcc8060063, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.