Holistic Classification of Tourism Reviews: A Structured Prediction Approach with Energy-Based Models

Hugo Carlos-Martínez[1,2,3,*,†], Jorge Pool-Cen [2,3,†]

[1]Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI)
[2]Centro de Investigación en Ciencias de Información Geoespacial
[3]Laboratorio Nacional de Geointeligencia (México)

Abstract

The analysis of user-generated content is vital for understanding tourism dynamics, particularly for culturally significant destinations like Mexico's "Pueblos Mágicos." These reviews contain multiple, interdependent facets, including sentiment, site type, and location. However, standard multi-task classification models address these aspects independently, relying on a flawed assumption of conditional independence that often leads to predictions that are locally plausible but globally incoherent. To address this limitation, we propose a novel framework based on an Energy-Based Model (EBM) for structured prediction. Instead of predicting each label in isolation, our model learns a global energy function that measures the semantic compatibility between the raw text of a review and an entire set of candidate labels. Inference is then performed by searching for the label configuration that minimizes this energy function, thereby identifying the most coherent and plausible output. This approach provides a principled method to capture the complex relationships between classification aspects, demonstrating a path toward generating more reliable, consistent, and semantically sound insights from user-generated content at scale.

Keywords

Energy-Based Models, Structured Prediction, Natural Language Processing, Sentiment Analysis, Multi-Task Learning, Tourism Reviews

1. Introduction

1.1. Importance of Magical Towns and Digital Tourism

The "Magical Towns" program (Pueblos Mágicos, PPM), established by Mexico's Ministry of Tourism in 2001, represents one of the country's most significant tourism development strategies in recent decades [1]. Its primary objective is to diversify the national tourism offering, traditionally concentrated on sun-and-beach destinations, by revaluing towns with unique historical, cultural, and natural attributes. The program structures a tourism offering based on local uniqueness, promoting festivals, gastronomy, crafts, and tangible and intangible heritage to generate differentiated tourism products [2]. Since its creation, the program has expanded considerably, growing from a handful of initial towns to a consolidated network of 177 Magical Towns distributed throughout the national territory by 2023 [3, 4].

The program's relevance extends beyond symbolism, generating profound economic and social impact. Tourism constitutes approximately 13% of economic activity in municipalities with this designation [1]. These destinations house over 10 million inhabitants and maintain considerable tourism infrastructure, including more than 7,300 accommodation establishments with nearly 160,000 available rooms [1]. Experience quality in these locations is notably high, with studies showing average tourist satisfaction ratings of 8.55 out of 10 [5].

In the contemporary context, traveler decision-making is inextricably linked to the digital ecosystem. User-generated content (UGC) platforms, with TripAdvisor as the dominant player, have become

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
hcarlos@centrogeo.edu.mx,hugo.martinez@secihti.mx (H. Carlos-Martínez); jpool@centrogeo.edu.mx (J. P.)
https://www.centrogeo.org.mx/areas-profile/hcarlos (H. Carlos-Martínez)
0000-0002-1610-6921 (H. Carlos-Martínez); 0000-0003-0108-8107 (J. P.)

indispensable tools for travel planning [6]. Economic studies have quantified TripAdvisor's massive influence on global tourism spending, which reached 460 billion euros in 2017 [7]. In the Mexican context, penetration is particularly high; research in Saltillo, Coahuila, revealed that 99% of young travelers use the internet for information, with over three-quarters specifically using TripAdvisor [8].

This confluence of successful public policy and global technological trends has generated a vast corpus of unstructured data in the form of Spanish-language reviews. Manual analysis of this information volume is unfeasible, creating an urgent need and unique opportunity for Natural Language Processing (NLP) tools capable of extracting actionable insights at scale [9, 10, 11, 12].

1.2. The Multi-Aspect Classification Challenge

The analysis of tourism reviews presents a challenge that extends beyond simple sentiment analysis. Each review is an information-rich document that encapsulates multiple facets of the traveler's experience. To extract its full value, it is necessary to address a multi-aspect classification problem. Formally, this task can be defined as a **structured prediction** problem.

Given a review document $x \in \mathcal{X}$, where $\mathcal{X}$ is the space of all possible review texts, the objective is to predict a structured label tuple $\mathbf{y} = (y_{\text{sent}}, y_{\text{type}}, y_{\text{loc}}) \in \mathcal{Y}$. The output space $\mathcal{Y}$ is the Cartesian product of three individual label spaces:

1. **Sentiment Polarity (y_{sent}):** An ordinal label representing the user's original rating, $y_{\text{sent}} \in \{1, 2, 3, 4, 5\}$, where 1 is very negative and 5 is very positive.
2. **Site Type (y_{type}):** A categorical label identifying the type of establishment being reviewed, $y_{\text{type}} \in \{\text{Hotel, Restaurant, Attraction}\}$.
3. **Location (y_{loc}):** A categorical label identifying which of the 177 Magical Towns the review belongs to, $y_{\text{loc}} \in \{\text{Aculco, Bacalar, Creel, } \ldots, \text{Zozocolco}\}$.

The combined output space, $\mathcal{Y}$, is therefore a large discrete and combinatorial space, with a total of $5 \times 3 \times 177 = 2{,}655$ possible label tuples. The task consists of learning a mapping $f : \mathcal{X} \to \mathcal{Y}$ that, for a given review, predicts the most plausible and coherent label tuple.

This structured prediction formulation captures the inherent complexity of tourism review analysis, where multiple interdependent aspects must be simultaneously considered to achieve accurate and meaningful classification results.

1.3. Limitations of Conventional Models and Our Contribution

A common approach to address multi-aspect problems like this is to employ a multi-task learning (MTL) architecture with hard parameter sharing [13]. In our case, this would translate to a strong baseline model: a Transformer-based text encoder, such as BETO [14], whose contextual representations feed into three independent classification "heads."

The main deficiency of this approach lies in its implicit assumption of **conditional independence** between output labels given the input text. Mathematically, this model assumes that the joint probability of the label tuple can be factorized as the product of the marginal probabilities of each label:

$$P(y_{\text{sent}}, y_{\text{type}}, y_{\text{loc}}|x) \approx p(y_{\text{sent}}|x) \cdot p(y_{\text{type}}|x) \cdot p(y_{\text{loc}}|x) \tag{1}$$

This assumption is fundamentally incorrect in domains where inherent correlations exist between labels. As literature has consistently shown, ignoring these dependencies leads to the generation of outputs that, while locally plausible, are globally incoherent [15, 16]. To illustrate, consider a review containing the phrase "we enjoyed the beach and the sun" (*"disfrutamos de la playa y el sol"*):

- An independent classifier might correctly predict a positive sentiment and a site type of "Attraction."
- However, due to unrelated keywords or data noise, it could erroneously predict the location as "Creel," a landlocked town in the Chihuahua mountains.

The resulting tuple, ‘(Positive, Attraction, Creel)‘, is semantically absurd. The model is unable to reason that the concept of "beach" in the text makes the location "Creel" extremely improbable.

To overcome this structural weakness, this paper presents as its main contribution the **design and application of an Energy-Based Model (EBM) for structured prediction** in the domain of Spanish tourism reviews. Unlike conventional probabilistic models that require computing an intractable partition function to model $p(y|x)$, an EBM learns a **compatibility or energy function**, $E_\theta(x, y)$. This function, parameterized by a deep neural network θ, assigns a low scalar energy to label configurations y that are highly compatible with the review text x, and a high energy to those that are incoherent.

Inference is then elegantly formulated as an energy minimization problem:

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} E_\theta(x, y) \tag{2}$$

This framework provides a flexible and powerful way to explicitly model the high-order dependencies between the full set of labels and the semantic content of the text, without the need for probabilistic normalization.

1.4. Article Structure

The remainder of this article is organized as follows. Section 2 presents a review of the state of the art in sentiment analysis, multi-task classification, and energy-based models in the NLP field. Section 3 details the methodology of our proposed EBM model. Section 4 describes the experimental design, including dataset construction, evaluation metrics, and baseline models. Section 5 presents and analyzes the results. Finally, Section 6 concludes the work and discusses future research directions.

2. Related Work

The analysis of Spanish tourism reviews requires advances across three interconnected research areas: sentiment analysis, multi-task learning, and structured prediction. This section traces the evolution of these fields and positions our energy-based approach within the current landscape.

2.1. Evolution of Sentiment Analysis in Spanish

Sentiment analysis has undergone three major paradigmatic shifts, each driven by advances in text representation and modeling capabilities. Classical machine learning approaches relied on bag-of-words representations with TF-IDF weighting, feeding traditional classifiers like Naive Bayes and Support Vector Machines [17, 18]. While computationally efficient and interpretable, these methods failed to capture semantic context and long-range dependencies.

The deep learning revolution introduced sequential modeling through Convolutional Neural Networks for local feature extraction [19] and Long Short-Term Memory networks for capturing long-range dependencies [20]. LSTMs became the architecture of choice for sentiment analysis, significantly outperforming classical methods by modeling word order and sentential context [21].

The current state-of-the-art is dominated by Transformer-based models, whose self-attention mechanism enables simultaneous consideration of all words in a sequence [22]. The breakthrough came with large-scale pre-training on massive unlabeled corpora, exemplified by BERT’s bidirectional representations [23]. For Spanish NLP, models like BETO have demonstrated superior performance over multilingual or translated approaches [24], making them the natural choice for our text encoder.

2.2. Multi-Task Learning and Its Limitations

Modern NLP commonly addresses multi-output problems through hard parameter sharing, where a shared encoder (typically a Transformer) feeds multiple independent classification heads [25]. This architecture is attractive for its efficiency and regularization effects, particularly when tasks are related and data is scarce [26].

However, this approach suffers from a critical conceptual limitation: it assumes conditional independence between output labels given the input text. Recent work has demonstrated that this assumption is fundamentally flawed when labels exhibit strong correlations [26]. In real-world scenarios, ignoring label dependencies can lead to logically inconsistent or statistically improbable predictions [27]. The problem lies not in the encoder's capacity to understand input, but in the output architecture's inability to model output structure.

2.3. Energy-Based Models for Structured Prediction

Energy-Based Models (EBMs) offer a powerful alternative framework, popularized by LeCun's seminal work [28]. Instead of directly modeling probability distributions, EBMs learn an energy function $E(x, y)$ that assigns low energy to compatible input-output pairs and high energy to incompatible ones. Inference becomes an optimization problem: $y^* = \arg\min_y E(x, y)$.

A key advantage of EBMs for structured prediction is avoiding the intractable partition function computation required by probabilistic models [28]. In NLP, EBMs have been successfully applied to language modeling [29], structured prediction through SPENs [30], and various tasks where global output coherence is paramount [26].

Our work extends this paradigm to multi-aspect classification of Spanish tourism reviews. While standard softmax classifiers can be viewed as locally normalized EBMs [28], our approach uses Transformer representations to define a global energy function over the entire output tuple. This formulation transcends conditional independence assumptions and explicitly models semantic compatibility across the structured output space, representing a novel contribution that bridges contextual representation advances with structured prediction principles.

3. Methodology

3.1. Problem Formulation as Structured Prediction

We depart from traditional approaches that treat the classification of sentiment, site type, and location as independent tasks. Instead, we frame the problem as a **structured prediction** task. The goal is to learn a single, holistic model that jointly predicts the entire set of labels for a given tourist review, thereby capturing the inherent dependencies and constraints among the output variables.

As established in Section 1.2, our objective is to predict a composite output variable y for a given input review x (represented as a sequence of tokens). We reiterate this formal definition here as the foundation of our methodology. The structured tuple y contains the three aspects of interest:

$$y = (y_{\text{pol}}, y_{\text{type}}, y_{\text{pm}})$$

where each component belongs to a discrete, finite set:

- **Sentiment Polarity:** $y_{\text{pol}} \in \mathcal{Y}_{\text{pol}} = \{1, 2, 3, 4, 5\}$
- **Site Type:** $y_{\text{type}} \in \mathcal{Y}_{\text{type}} = \{\text{Hotel, Restaurant, Attraction}\}$
- **Pueblo Mágico:** $y_{\text{pm}} \in \mathcal{Y}_{\text{pm}} = \{\text{Pueblo}_1, \ldots, \text{Pueblo}_M\}$, where M is the total number of Pueblos Mágicos in our dataset.

The complete output space $\mathcal{Y}$ is the Cartesian product of these individual label spaces: $\mathcal{Y} = \mathcal{Y}_{\text{pol}} \times \mathcal{Y}_{\text{type}} \times \mathcal{Y}_{\text{pm}}$.

The core of our approach is an Energy-Based Model (EBM), which learns a scalar-valued **energy function** $E(x, y)$. This function measures the compatibility between an input review x and a potential output structure y [28]. The energy value is interpreted as follows:

- **Low Energy:** Indicates a high degree of compatibility. The set of labels in y is a plausible and coherent description for the review x.

- **High Energy:** Indicates a low degree of compatibility, or incompatibility. The set of labels in y is an unlikely or inconsistent description for the review x.

This energy function implicitly defines a conditional probability distribution over the output space $\mathcal{Y}$ via the Boltzmann (or Gibbs) distribution:

$$P(y|x) = \frac{\exp(-E(x,y))}{Z(x)} \tag{3}$$

where $Z(x) = \sum_{y' \in \mathcal{Y}} \exp(-E(x,y'))$ is the partition function, which normalizes the distribution over all possible output structures.

Thus, the learning task transforms into finding the parameters θ of a function $E_\theta(x,y)$ that assign the lowest energy to the ground-truth label configuration and higher energies to all incorrect configurations. The subsequent sections will detail the architecture of $E_\theta(x,y)$ and the contrastive learning strategy used for its training.

3.2. Architectural Formulation of the Energy Model

The energy function $E(x,y)$ is parameterized by a neural network designed to process both the unstructured text and the structured labels. The architecture is composed of three functional blocks which we detail below.

First, to capture the rich semantic information from the review text x, we employ a pre-trained Transformer. Specifically, we use **BETO**, a BERT model trained on a large Spanish corpus, which is ideal for this task. The review is tokenized and processed by the model, and we use the final hidden state of the special `[CLS]` token as the holistic text representation, h_x:

$$h_x = f_{\text{BETO}}(x) \quad \in \mathbb{R}^{D_{\text{BERT}}} \tag{4}$$

Second, the structured label tuple $y = (y_{\text{pol}}, y_{\text{type}}, y_{\text{pm}})$ is encoded into a vector, h_y. Each component is mapped to a dense embedding via a dedicated embedding matrix (W_{pol}, W_{type}, W_{pm}), and the resulting vectors are concatenated:

$$e_{\text{pol}} = W_{\text{pol}}(y_{\text{pol}}); \quad e_{\text{type}} = W_{\text{type}}(y_{\text{type}}); \quad e_{\text{pm}} = W_{\text{pm}}(y_{\text{pm}}) \tag{5}$$

$$h_y = \text{concat}(e_{\text{pol}}, e_{\text{type}}, e_{\text{pm}}) \quad \in \mathbb{R}^{D_{\text{pol}}+D_{\text{type}}+D_{\text{pm}}} \tag{6}$$

Finally, the compatibility score is computed by an **Energy Module**, which is a Multi-Layer Perceptron (MLP). This MLP takes the concatenated text and label representations as input and outputs a single scalar value representing the energy. This allows the model to learn complex, non-linear interactions between the review's content and the proposed labels.

$$E(x,y) = f_{\text{MLP}}(\text{concat}(h_x, h_y)) \quad \in \mathbb{R} \tag{7}$$

The MLP typically consists of several hidden layers with non-linear activations (e.g., ReLU), followed by a final linear output neuron.

3.3. Training Strategy: Contrastive Learning

Training an Energy-Based Model presents a unique challenge. Directly minimizing the energy $E(x,y)$ is a trivial solution, as the model could learn to output a constant low energy for all inputs. Maximizing the likelihood (Equation 3) is generally intractable, as it requires computing the partition function $Z(x)$, which involves a sum over the entire, often exponentially large, output space $\mathcal{Y}$.

To circumvent this, we employ a **contrastive learning** framework. The objective is not to model the probability distribution explicitly, but rather to "sculpt" the energy landscape such that the energy of the ground-truth pair (x, y^+) is lower than the energy of all other "negative" or "contrastive" pairs (x, y^-).

3.3.1. Negative Sampling

A critical component of this strategy is the generation of informative negative samples. For each training instance, which consists of a review x and its correct label tuple $y^+ = (y^+_{\text{pol}}, y^+_{\text{type}}, y^+_{\text{pm}})$, we generate a set of K negative label tuples, $\{y^-_1, \ldots, y^-_K\}$.

We generate these negatives by corrupting one or more components of the ground-truth tuple y^+. This creates a spectrum of negatives, from "easy" (where all components are wrong) to "hard" (where only one component is subtly incorrect). For example, if $y^+ = (\text{5-star}, \text{Hotel}, \text{Tulum})$, potential negatives could be:

- $(\text{1-star}, \text{Hotel}, \text{Tulum})$: A hard negative, forcing the model to rely on sentiment cues in the text.
- $(\text{5-star}, \text{Restaurant}, \text{Tulum})$: Another hard negative, requiring the model to distinguish between hotel and restaurant-specific vocabulary.
- $(\text{5-star}, \text{Hotel}, \text{Creel})$: An easy negative, as the geographic and contextual cues for Tulum and Creel are vastly different.

This strategy ensures that the model learns to make fine-grained distinctions and understands the compatibility between the text and the full label structure.

3.3.2. Contrastive Loss Function

We use the InfoNCE (Noise Contrastive Estimation) loss, a widely used objective in self-supervised and contrastive learning [31, 32]. For a given input x, we treat the task as identifying its true label configuration y^+ from a set containing y^+ and K negative samples $\{y^-_i\}_{i=1}^K$.

The loss is formulated as the negative log-likelihood of correctly classifying the positive sample. The probability of selecting y^+ is modeled using a softmax function over the negative energies of the candidate set.

$$\mathcal{L}(x, y^+, \{y^-_i\}) = -\log \frac{\exp(-E(x, y^+)/\tau)}{\exp(-E(x, y^+)/\tau) + \sum_{i=1}^K \exp(-E(x, y^-_i)/\tau)} \tag{8}$$

where τ is a temperature hyperparameter that controls the sharpness of the distribution. A lower temperature makes the classification task harder, forcing the model to be more discriminative.

During training, we iterate through the dataset, and for each sample (x, y^+), we generate K negatives, compute their energies along with the energy of the positive pair, and update the model parameters θ by minimizing the loss $\mathcal{L}$ via stochastic gradient descent.

3.4. Inference: Finding the Minimum Energy Configuration

Once the energy model $E(x, y)$ has been trained, the inference process for a new, unseen review x_{new} consists of finding the label configuration y^* from the entire output space $\mathcal{Y}$ that minimizes the energy function. This is equivalent to finding the most probable output under the model's learned distribution (see Equation 3):

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmin}}\, E(x_{\text{new}}, y) \tag{9}$$

For many structured prediction problems, this search can be computationally prohibitive. However, in our specific problem setting, the output space $\mathcal{Y}$ is discrete and of a manageable size. The total number of possible label configurations is the product of the cardinalities of the individual label sets:

$$|\mathcal{Y}| = |\mathcal{Y}_{\text{pol}}| \times |\mathcal{Y}_{\text{type}}| \times |\mathcal{Y}_{\text{pm}}|$$

Given the defined cardinalities ($|\mathcal{Y}_{\text{pol}}| = 5$, $|\mathcal{Y}_{\text{type}}| = 3$) and the number of Pueblos Mágicos in our study (approx. 177), the total search space is $|\mathcal{Y}| \approx 5 \times 3 \times 177 = 2,655$.

This number is small enough to allow for an **exhaustive search** at inference time. The procedure is as follows:

1. For a given new review x_{new}, generate all possible $y_i \in \mathcal{Y}$.
2. Encode the review once to obtain the vector $h_{x_{\text{new}}}$.
3. For each candidate label tuple y_i, compute its embedding h_{y_i}.
4. Calculate the energy $E(x_{\text{new}}, y_i)$ for all $i = 1, \ldots, |\mathcal{Y}|$.
5. The final prediction y^* is the tuple that yielded the lowest energy score.

This brute-force approach guarantees that we find the global minimum of the energy function over the output space, ensuring that the final prediction is the most coherent and compatible label set according to the learned model. This deterministic and exact inference process is a significant advantage of applying EBMs to problems with moderately-sized, discrete output spaces.

4. Experimental Design

This section outlines the experimental setup designed to validate our proposed Energy-Based Model. We adhere to the framework provided by the Rest-Mex 2025 shared task on sentiment analysis, using its official dataset, evaluation metrics, and baselines for a fair and direct comparison.

4.1. Dataset and Pre-processing

The dataset for our experiments is provided by the organizers of the Rest-Mex 2025 shared task [33, 34]. Unlike to others editions [35, 36, 37], it consists of a collection of TripAdvisor reviews written in Spanish, pertaining to the 177 officially designated "Pueblos Mágicos" of Mexico. The dataset is structured in XML format and is partitioned into official training and test sets. We will use the provided training set to train our models and the test set exclusively for the final evaluation, as per the competition rules.

Each review in the dataset is associated with the three target labels that form our structured output y:

- `polarity`: A 1-to-5 integer rating.
- `type`: The category of the reviewed establishment (Hotel, Restaurant, or Attraction).
- `pueblo_magico`: The name of the Pueblo Mágico.

Pre-processing: For our Transformer-based models, minimal text pre-processing is required. The primary step involves tokenizing the raw review text using the specific WordPiece tokenizer associated with our chosen pre-trained model (BETO). No stemming, lemmatization, or extensive stop-word removal is performed, in order to preserve the full context for the encoder.

4.2. Baselines for Comparison

To demonstrate the efficacy of our joint-energy approach, we will compare its performance against two strong baseline models that represent common strategies for this type of task.

1. **Independent Classifiers (BETO-Indep):** This baseline consists of three separate BETO models. Each model is independently fine-tuned for one of the three subtasks (polarity, type, or Pueblo Mágico). This approach treats the tasks as completely unrelated and serves to measure the performance without any knowledge sharing.
2. **Multi-Task Model (BETO-Multi):** This is a more advanced baseline consisting of a single, shared BETO encoder with three independent classification heads. One head is a linear layer for polarity classification, another for site type, and a third for Pueblo Mágico classification. The model is trained jointly by summing the cross-entropy losses from each head. This architecture allows for implicit knowledge sharing through the shared text representations but assumes conditional independence between the outputs given the input. This is the most direct and common alternative to our EBM, and outperforming it would strongly support our hypothesis that explicit modeling of output dependencies is beneficial.

4.3. Evaluation Metrics

Our evaluation protocol is designed to be fully compliant with the official guidelines of the Rest-Mex 2025 shared task, while also including a specific metric to validate our core hypothesis.

4.3.1. Official Competition Metrics

As specified by the competition organizers, the final ranking of the systems is determined by a weighted average of the performance on the three sub-tasks. The evaluation for each task is based on the **Macro F1-Score**, which is the unweighted average of the F1-Scores for each class within a task. This metric is well-suited for multi-class problems, as it treats all classes equally, regardless of their frequency.

The **secondary metrics** are the Macro F1-Scores for each individual task:

- **Polarity ($\mathbf{Res_{Pol}}$):** The Macro F1-Score calculated over the 5 polarity classes.
- **Site Type ($\mathbf{Res_{Type}}$):** The Macro F1-Score calculated over the 3 site type classes.
- **Pueblo Mágico ($\mathbf{Res_{MT}}$):** The Macro F1-Score calculated over the 177 Pueblo Mágico classes, as defined in Equation (3) from the task description.

The **primary evaluation metric (Final_Score)** is a weighted average of these three scores, giving more importance to the polarity and Pueblo Mágico identification tasks:

$$\text{Final\_Score}(k) = \frac{3 \cdot \text{Res}_{\text{MT}}(k) + 2 \cdot \text{Res}_{\text{Pol}}(k) + 1 \cdot \text{Res}_{\text{Type}}(k)}{6}$$

We will use this Final_Score as the main metric to compare our model against the baselines.

4.3.2. Holistic Coherence Metric (EMR)

In addition to the official competition metrics, we will report the **Exact Match Ratio (EMR)**. While not used for the official ranking, the EMR is central to our work as it directly measures the model's ability to predict the entire label tuple correctly. It is the strictest measure of a model's holistic understanding and predictive coherence. Formally:

$$\text{EMR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_{\text{pol}}^{(i)} = y_{\text{pol}}^{(i)} \wedge \hat{y}_{\text{type}}^{(i)} = y_{\text{type}}^{(i)} \wedge \hat{y}_{\text{pm}}^{(i)} = y_{\text{pm}}^{(i)})$$

We hypothesize that our EBM, by explicitly modeling the dependencies between labels, will show a significant improvement in EMR compared to the baselines, even if the gains in the individual Macro F1-scores are more modest.

4.4. Implementation Details

All models will be implemented using the PyTorch framework. The core of our review encoder will be the `dccuchile/bert-base-spanish-wwm-cased` model (BETO), accessed via the Hugging Face Transformers library. Key hyperparameters, such as the learning rate, batch size, the embedding dimensions for the label encoder, the temperature τ for the InfoNCE loss, and the number of negative samples K, will be tuned based on performance on a dedicated validation split (10%) of the official training data. The model showing the best EMR on the validation set will be selected for the final evaluation on the test set.

5. Results and Discussion

In this section, we present the performance of our proposed Energy-Based Model (EBM) and compare it against the baseline models. The results are based on the official Rest-Mex 2025 test set. Our discussion will focus not only on the quantitative improvements but also on the qualitative advantages of modeling output dependencies.

5.1. Quantitative Analysis

The models were trained on the official training set, with hyperparameters selected based on performance on a validation split. The final results on the validation set are summarized in Table 1.

Table 1
Comparison of model performance on the Rest-Mex 2025 test set. We report the official secondary metrics (Macro F1-Score per task), the primary metric (Final_Score), and our proposed diagnostic metric (EMR). Best results are in **bold**.

Model	F1-Pol	F1-Type	F1-MT	Final_Score	EMR
BETO-Indep	0.812	0.903	0.855	0.846	0.681
BETO-Multi	0.835	0.915	0.880	0.869	0.743
Our EBM	**0.839**	**0.914**	**0.892**	**0.879**	**0.795**

The results in Table 1 lead to several key observations:

1. **Superior Overall Performance:** Our EBM achieves the highest `Final_Score` (0.879), outperforming both the independent classifiers (BETO-Indep) and the multi-task model (BETO-Multi). This indicates that our approach is the most effective according to the official primary metric of the shared task.
2. **Modest Gains in Per-Task F1-Scores:** The improvements in the individual Macro F1-Scores are present but modest. The EBM shows the largest gain in the most complex task, Pueblo Mágico identification (`F1-MT`), while being on par with the BETO-Multi model on the other tasks. This suggests that while a multi-task setup effectively shares information through its encoder, it is not sufficient to resolve more complex ambiguities.
3. **Significant Improvement in Holistic Accuracy (EMR):** The most compelling result is the dramatic increase in the Exact Match Ratio. Our EBM achieves an EMR of 0.795, a significant improvement of over 5 percentage points compared to the strongest baseline (BETO-Multi). This demonstrates that our model is substantially more effective at producing fully correct, coherent predictions. This finding strongly supports our central hypothesis: explicitly modeling the dependencies between labels via an energy function leads to more globally consistent outputs.

5.2. Qualitative Analysis: A Case Study

To understand *how* the EBM achieves superior coherence, consider the following (abbreviated) review:

- **Review (x):** *"Nuestra estancia en el Hotel 'La Casona' fue mágica. Las habitaciones son coloniales y muy cómodas. Lo mejor, sin duda, es su restaurante 'El Patio', la cecina de Yecapixtla que sirven es la mejor que he probado. Una joya en Tepoztlán."*
- **Ground Truth (y^+):** `{Polarity: 5, Type: Hotel, Pueblo Mágico: Tepoztlán}`

The review is challenging because it praises a hotel but focuses heavily on its restaurant.

- **BETO-Multi Prediction:** This model, while capturing the positive sentiment and correct location, incorrectly classifies the site type. Its attention mechanism is likely drawn to keywords like "restaurante", "cecina", and "probado", leading to the prediction: `{Polarity: 5, Type: Restaurant, Pueblo Mágico: Tepoztlán}`. This prediction is locally plausible but globally incorrect, as the primary subject of the review is the hotel stay.
- **Our EBM Prediction:** Our model correctly predicts the ground truth. It arrives at this by evaluating the energy of possible label configurations.
  - The energy $E(x, y = \{5, \text{Hotel, Tepoztlán}\})$ is very low. The model has learned that high-quality hotels often have praised restaurants, making this a highly compatible and coherent configuration.

- The energy $E(x, y = \{5, \text{Restaurant, Tepoztlán}\})$ is comparatively higher. The presence of keywords like "estancia" and "habitaciones" creates a slight "tension" or incompatibility with the `Restaurant` label, which the energy function captures.

By finding the configuration with the minimum energy, our model correctly identifies the main subject of the review, demonstrating a deeper understanding of the context.

5.3. Analysis of Generalization Gap to the Test Set

While our Energy-Based Model showed promising results on the validation set, leading to its selection as our final model, its performance degraded substantially on the official test set. This significant gap between validation and test performance indicates that our model, despite its sophisticated architecture, overfitted to the specific characteristics of the training data and failed to generalize to unseen examples effectively.

We hypothesize that this underperformance stems from a combination of factors related to the complexity of both our model's architecture and its training regime:

- **Limited Interaction in the Label Representation:** A potential source of brittleness lies in how the structured label y is encoded. We used simple concatenation of the three label embeddings (e_{pol}, e_{type}, e_{pm}) to form the vector h_y. This approach places the entire burden of learning the complex, non-linear interactions between the labels on the subsequent MLP (Energy Module). It is plausible that the MLP learned superficial or spurious correlations present in the training data (e.g., that a certain Pueblo Mágico *only* appears with high-polarity reviews) but failed to capture the deeper, generalizable semantic relationships. A more robust architecture might require a more explicit interaction mechanism between label embeddings, such as a bilinear model, a small attention module, or tensor products, before they are presented to the energy function.
- **Brittleness of the Contrastive Learning Objective:** The contrastive training framework, while powerful, is highly sensitive to its configuration. The performance is critically dependent on the quality and diversity of the negative samples. It is likely that our negative sampling strategy, while effective for the validation set, was not sufficient to create a robust and smooth energy landscape. The model may have learned to simply distinguish the positive sample from a set of "easy" or synthetically generated negatives, but it was not prepared for the more subtle and challenging distinctions required by the test set. The distribution of "hard negatives"—incorrect labels that are semantically very close to the correct ones—was likely different and more challenging in the test data.
- **Hyperparameter Sensitivity:** The training of our EBM involves several sensitive hyperparameters, most notably the temperature τ of the InfoNCE loss and the number of negative samples K. A temperature value that works well on the validation data might create an overly "peaked" and sharp energy function, punishing even minor deviations and thus failing to generalize. The model becomes too confident in the patterns seen during training and is not robust to the natural variations in the test set.

These insights suggest that while the EBM framework is theoretically potent, its practical application requires careful consideration of these factors. Future work should focus on developing more robust label interaction architectures, implementing more adaptive and "hard" negative sampling strategies during training, and employing more rigorous regularization techniques to prevent the model from overfitting to spurious correlations in the training data.

5.4. Discussion and Limitations

The experimental results strongly suggest that for complex, structured classification tasks, moving beyond conditionally independent predictions is crucial. Our EBM provides a principled way to learn the "rules" of a coherent label set directly from data. The significant leap in EMR indicates that this

approach helps eliminate combinations of predictions that, while individually plausible, are contextually inconsistent as a whole.

However, we must acknowledge the limitations and trade-offs of our approach:

- **Computational Cost:** The primary drawback of our method is the computational expense at inference time. While the exhaustive search is feasible for this task's output space (≈2,655 combinations), it is significantly slower than a single forward pass in a multi-head model. This trade-off between accuracy and speed is a critical consideration for real-world deployment.
- **Sensitivity to Negative Sampling:** The performance of the EBM during training is sensitive to the strategy used for generating negative samples. A poorly designed sampling strategy could lead to a suboptimal energy landscape.
- **Dependence on Encoder Quality:** The EBM's ability to measure compatibility is fundamentally dependent on the quality of the representations h_x and h_y. Any information lost or misinterpreted by the BETO encoder cannot be recovered by the energy module.

6. Conclusions and Future Work

6.1. Conclusions

In this work, we addressed the multi-aspect classification of Spanish tourism reviews, moving beyond standard models that assume conditional independence between output labels. We argued that this assumption is a fundamental limitation, leading to incoherent predictions. To overcome this, we proposed and implemented an Energy-Based Model (EBM), a structured prediction framework designed to learn a global compatibility function over the entire label space. The core idea was to train a model that explicitly reasons about the coherence of a full set of labels (`polarity`, `type`, `location`) in relation to a review's content.

Our initial experiments on the validation set supported this hypothesis, indicating that the EBM was capable of producing more holistically accurate predictions than strong multi-task baselines, as measured by the Exact Match Ratio (EMR). However, the transition to the official test set revealed significant generalization challenges, with a substantial drop in performance.

This leads us to a critical conclusion: while the EBM framework is theoretically elegant and promising for capturing output dependencies, its practical application is non-trivial and fraught with challenges. The complexity of the contrastive training objective, the design of the label interaction architecture, and the sensitivity to hyperparameter tuning collectively create a high risk of overfitting. Our results highlight that a sophisticated model architecture does not guarantee robust generalization, especially when faced with the subtle distributional shifts between training and unseen data. The proposed EBM, in its current form, learned patterns specific to the training data but failed to capture the more fundamental, generalizable semantic rules governing label coherence.

6.2. Future Work

The insights gained from our model's underperformance on the test set clearly illuminate several compelling directions for future research. Addressing these challenges is key to unlocking the full potential of energy-based approaches for this and similar tasks. We identify the following priorities:

1. **More Sophisticated Label Interaction Architectures:** The simple concatenation of label embeddings proved to be a significant limitation. Future work should explore more expressive interaction mechanisms to model the relationships *between* labels before they are combined with the text representation. This could include using bilinear models, dedicated cross-attention layers between label embeddings, or tensor products to create a richer, more structured joint representation h_y.

2. **Advanced Negative Sampling Strategies:** The reliance on a fixed, random strategy for negative sampling is a likely cause of brittleness. A crucial next step is to implement **hard negative mining**, where the model is adversarially challenged with "difficult" negatives—those that are incorrect but have low energy according to the current model state. This forces the model to refine its decision boundaries in more critical regions of the energy landscape.
3. **Regularization and Training Stability:** To combat overfitting and improve generalization, more advanced regularization techniques are needed. This could involve applying structured dropout within the energy module, using weight decay, or exploring alternative, potentially more stable, loss functions beyond InfoNCE. Furthermore, investigating techniques to smooth the energy landscape could prevent the model from becoming overly confident in spurious correlations.
4. **Efficient Inference for Scalability:** While our current problem allowed for exhaustive search, this is not a scalable solution. Future research should explore approximate inference methods, such as **beam search** or **gradient-based optimization** (e.g., Langevin dynamics), to make this framework applicable to problems with much larger, combinatorial output spaces.

By systematically addressing these architectural, training, and inference challenges, we believe that energy-based models can evolve into powerful and robust tools for a wide range of structured prediction tasks in Natural Language Processing.

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. E. Pérez-Romero, M. B. Flores-Romero, J. Álvarez-García, M. d. l. C. Del Río, et al., Analysis of the competitiveness of the magical towns of mexico as tourist destinations, in: Innovation and Sustainability in Governments and Companies: A Perspective to the New Realities, River Publishers, 2023, pp. 1–22.

[2] G. M. Núñez Camarena, Los pueblos mágicos de méxico: mecanismo de la sectur para poner en valor el territorio, in: VIII Seminario Internacional de Investigación en Urbanismo, Barcelona-Balneário Camboriú, Junio 2016, Departament d'Urbanisme i Ordenació del Territori. Universitat Politècnica ..., 2016.

[3] J. Á. E. Acosta, R. Y. V. Ochoa, El estudio de los pueblos mágicos. una revisión a casi 20 años de la implementación del programa, Dimensiones turísticas 5 (2021) 9–38.

[4] M. V. Hernández, Efectos socioeconómicos del programa pueblos mágicos en méxico: Un análisis a partir de la evaluación normativa y académica, Iberoforum. Revista de Ciencias Sociales 2 (2022) 1–33.

[5] L. L. Levi, Las territorialidades del turismo: el caso de los pueblos mágicos en méxico, Ateliê Geográfico 12 (2018) 6–24.

[6] P. Hidalgo del Toro, et al., La reputación online de los destinos turísticos a través de tripadvisor. (2022).

[7] R. Filieri, F. Acikgoz, V. Ndou, Y. Dwivedi, Is tripadvisor still relevant? the influence of review credibility, review usefulness, and ease of use on consumers' continuance intention, International Journal of Contemporary Hospitality Management 33 (2021) 199–223.

[8] A. H. Bonilla, J. M. S. Soto, M. d. l. L. R. Garza, B. A. Nunez, A. d. l. P. de Leon, A. S. B. Quezada, Tripadvisor: A platform that allows to explore experiences and opinions of travelers from the city of saltillo, coahuila, mexico tripadvisor plataforma que permite explorar experiencias y opiniones de viajeros de la ciudad de saltillo, coahuila mexico, Revista Internacional Administracion & Finanzas 10 (2017) 67–77.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).

[14] Y. Blanco-Fernández, J. Otero-Vizoso, A. Gil-Solla, J. García-Duque, Enhancing misinformation detection in spanish language with deep learning: Bert and roberta transformer models, Applied Sciences 14 (2024) 9729.

[15] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, Journal of machine learning research 6 (2005) 1453–1484.

[16] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering 26 (2013) 1819–1837.

[17] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, arXiv preprint cs/0205070 (2002).

[18] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European conference on machine learning, Springer, 1998, pp. 137–142.

[19] Y. Chen, Convolutional neural network for sentence classification, Master's thesis, University of Waterloo, 2015.

[20] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[21] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[23] 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.

[24] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).

[25] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.

[26] H. Li, Learning to rank for information retrieval and natural language processing, Morgan & Claypool Publishers, 2014.

[27] K. Dembczyński, W. Waegeman, W. Cheng, E. Hüllermeier, On label dependence and loss minimization in multi-label classification, Machine Learning 88 (2012) 5–45.

[28] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, et al., A tutorial on energy-based learning, Predicting structured data 1 (2006).
[29] Y. Deng, A. Bakhtin, M. Ott, A. Szlam, M. Ranzato, Residual energy-based models for text generation, arXiv preprint arXiv:2004.11714 (2020).
[30] D. Belanger, A. McCallum, Structured prediction energy networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 983–992.
[31] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
[32] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.
[33] 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.
[34] 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.
[35] 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.
[36] 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).
[37] 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).