Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 425-436 recibido 30-06-2023 revisado 07-07-2023 aceptado 16-07-2023

Overview of Rest-Mex at IberLEF 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts

Resumen de la tarea Rest-Mex en IberLEF 2023: Investigación sobre Análisis de Sentimiento para Textos Turísticos Mexicanos

Miguel Á. Álvarez-Carmona,[1,2] **Ángel Díaz-Pacheco,**[4] **Ramón Aranda,**[1,2]
Ansel Y. Rodríguez-González,[1,3] **Victor Muñiz-Sánchez**,[1] **A. Pastor López-Monroy**,[1]
Fernando Sánchez-Vega,[1,2] **Lázaro Bustio-Martínez**[5]
[1]Centro de Invetsigación en Matemáticas
[2]Consejo Nacional de Humanidades, Ciencias y Tecnologías
[3]Centro de Investigación Científica y de Educación Superior de Ensenada
[4]Universidad de Guanajuato
[5]Universidad Iberoamericana
{miguel.alvarez, arac, victor_m, pastor.lopez, fernando.sanchez}@cimat.mx
angel.diaz@ugto.mx , ansel@cicese.edu.mx, lazaro.bustio@ibero.mx

Abstract: This paper presents the framework and results of the Rest-Mex task at IberLEF 2023, focusing on sentiment analysis and text clustering of tourist texts. The study primarily focuses on texts related to tourist destinations in Mexico, although this edition included data from Cuba and Colombia for the first time. The sentiment analysis task aims to predict the polarity of opinions expressed by tourists, classifying the type of place visited, whether it's a tourist attraction, hotel, or restaurant, as well as the country it is located in. On the other hand, the text clustering task aims to classify news articles related to tourism in Mexico. For both tasks, corpora were built using Spanish opinions extracted from TripAdvisor and news articles from Mexican media. This article compares and discusses the results obtained by the participants in both sub-tasks. Additionally, a method is proposed to measure the easiness of a multi-class text classification corpus, along with an approach for system selection in a possible late fusion scheme.
Keywords: Rest-Mex 2023, Sentiment Analysis, Clustering, Mexican Tourist Text.

Resumen: Este artículo presenta el marco y los resultados de la tarea Rest-Mex en IberLEF 2023, que se enfoca en el análisis de sentimiento y agrupamiento de textos turísticos. El estudio se centra principalmente en textos relacionados con destinos turísticos en México, aunque esta edición incluyó datos de Cuba y Colombia por primera vez. La tarea de análisis de sentimiento tiene como objetivo predecir la polaridad de opiniones expresadas por turistas, clasificando el tipo de lugar visitado, ya sea un atractivo turístico, un hotel o un restaurante, así como el país en el que se encuentra. Por otro lado, la tarea de agrupamiento de textos busca clasificar noticias relacionadas con el turismo en México. Para ambas tareas, se construyeron corpus utilizando opiniones en español extraídas de TripAdvisor y noticias de medios mexicanos. En este artículo, se comparan y discuten los resultados obtenidos por los participantes en ambas sub tareas. Además, se propone un método para medir la facilidad de un corpus de clasificación textual multi-clase, así como un enfoque para la selección de sistemas en un posible esquema de fusión tardía
Palabras clave: Rest-Mex 2023, Análisis de sentimientos, Agrupación de textos, Textos Turísticos.

ISSN 1135-5948 DOI 10.26342/2023-71-34

1 Introduction

Tourism is a multifaceted phenomenon encompassing social, cultural, and economic aspects, involving people's movement to destinations beyond their usual place of residence for personal or business/professional purposes (Guerrero-Rodriguez et al., 2021). This activity plays a crucial role in numerous countries (Diaz-Pacheco et al., 2022), including Mexico, where it ranks among the top ten globally and second among Iberoamerican countries in terms of international tourist arrivals (Álvarez-Carmona et al., 2022b)[1]. Moreover, tourism contributes significantly to Mexico's national GDP, accounting for 8.7% (Arce-Cardenas et al., 2021) and generating approximately 4.5 million direct jobs (Álvarez-Carmona et al., 2022a).

In 2021, Rest-Mex emerged as an evaluation forum focusing on text analysis within the domain of tourism, aiming to provide solutions for various tasks in Mexican Spanish (Álvarez-Carmona et al., 2021). The forum's inaugural edition featured two tasks: analysis of recommendation systems and sentiment analysis, both based on data collected from TripAdvisor.

Building upon its previous success, the 2022 edition of Rest-Mex expanded its scope by introducing three sub-tasks: Recommendation System, Sentiment Analysis on Mexican tourist texts, and a novel task involving the determination of the color of the Mexican Covid-19 epidemiological semaphore (Álvarez-Carmona et al., 2022c).

Continuing its commitment to advancing research in the field, the 2023 edition of Rest-Mex proposes two sub-tasks: *i*) Sentiment Analysis and *ii*) Clustering for Mexican tourism. To support these tasks, two distinct datasets have been curated. For sentiment analysis, a collection of **359,565** opinions was gathered from various tourist destinations in Mexico, Cuba, and Colombia. The data includes labeled information about polarity, type of attraction, and the country of origin for each opinion. Additionally, **114,550** news items related to tourism topics such as insecurity, gastronomy, prices, and landscapes, were collected from multiple states within the Mexican Republic.

[1]Mexico is in the world's top ten and the second Iberoamerican country related to the arrival of international tourists (Olmos-Martínez et al., 2023).

This study not only presents the collections and provides an overview of the participating works in the forum but also introduces two methodologies. The first methodology measures the easiness of the corpus, while the second methodology selects systems that make substantial contributions to the classification of all instances in the test set. Both approaches are founded on set operations.

This paper is structured as follows: Section 2 presents an overview of the collection-building process and the evaluation metrics used in the forum. Section 3 provides a summary of the solutions submitted by participants for the tasks, along with the analysis of their results. Finally, Section 4 presents the conclusions drawn from this evaluation forum's findings.

2 Evaluation framework

This section outlines the construction of the two used corpora, highlighting particular properties, challenges, and novelties. It also presents the evaluation measures used for the tasks.

2.1 Sentiment Analysis corpus

The first subtask is a classification task where the participating system can predict the polarity, the tourist attraction, and the country of an opinion issued by a tourist who traveled to the representative tourist places. This collection was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2023. Each opinion's polarity is an integer between [1, 5], where{1: Very bad, 2: Bad, 3: Neutral, 4: Good,5: Very good}. Also, the participants must determine the attractiveness of the opinion being issued. The possible classes are Attractive, Hotel, and Restaurant. Finally, through each opinion, the country that the tourist visited must also be determined. For this corpus, we have data from 3 countries: Mexico, Cuba, and Colombia.

The corpus consists of **359,565 opinions** shared by tourists. We use a 70/30 partition to divide into train and test. This means that we used 251,702 labeled instances for the train partition, while we used 107,863 unlabeled instances for the test partition.

Table 1 shows the distribution of the instances for the sentiment analysis task for the train and test partitions for polarity, type, and country.

Trait	Class	Train	Test
Polarity	1	5772	2560
	2	6952	2866
	3	21656	9133
	4	60227	25938
	5	157095	67366
	$\sum$	251702	107863
Type	Attractive	111188	47819
	Hotel	76042	32538
	Restaurant	64472	27506
	$\sum$	251702	107863
Country	Mexico	118776	50917
	Cuba	66223	28183
	Colombia	66703	28763
	$\sum$	251702	107863

Table 1: Instances distribution on sentiment analysis task.

The class imbalance (**?**) is clear for the 3 traits, making this a task with a significant degree of difficulty too.

Formally the problem of this task is defined as:

"Given an opinion about a tourist place, the goal is to determine the polarity, between 1 and 5, of the text, the visited attraction, which could be an attraction, a hotel, or a restaurant and the country visited among Mexico, Cuba, and Colombia."

2.2 Clustering for Mexican Tourist Texts

The second subtask is a clustering task where the participating system can group news containing information relevant to tourism. This collection was obtained from news websites that published reports regarding tourism from 2020 to December 2023. The collection has 4 important groups[2]: insecurity, gastronomy, prices, and landscapes.

For this task, **114,550** news items referring to tourism were collected for several states of the Mexican Republic. Unlike previous editions of Rest-Mex, this particular subtask is unique in that it is unsupervised. In contrast to other subtasks that involved a training phase, this subtask does not require one. Instead, all instances are utilized for comprehensive clustering.

[2]The participants did not know the name of these 4 groups

Group	Instances
Insecurity	96,872
Gastronomy	8,876
Prices	5,238
Landscapes	3,564
$\sum$	114,550

Table 2: Instances distribution for the clustering task.

Table 2 shows the distribution of the instances.

Similar to the sentiment analysis task, this corpus displays a pronounced class imbalance. This imbalance represents a significant challenge in achieving the desired groupings.

Formally the problem of this task is defined as:

"Given a tourist text news corpus C, each system must group the text in 4 groups"

2.3 Performance measures

Previous Rest-Mex editions have used the well-known MAE (Mean Average Error) metric. Nevertheless, for this edition, we propose to give more weight to minority classes. For the sentiment analysis collection of the Rest-Mex, the minority classes are the ones with the most negative polarities. Therefore, for this edition, to evaluate the result of the polarity classification, Equation 1 is proposed.

$$Res_P(k) = \frac{\sum_{i=1}^{|C|}((1 - \frac{T_{C_i}}{T_C}) * F_i(k))}{\sum_{i=1}^{|C|} 1 - \frac{T_{C_i}}{T_C}} \quad (1)$$

Where k is a forum participant system, $C = \{1, 2, 3, 4, 5\}$, T_C is the total instances in the collection, T_{C_i} is the total is instances in the class i. Finally, $F_i(k)$ is the F-measure value for the class i obtained by the system k. With this measure, correctly classified instances of class 1 will have more importance than instances of class 2, which in turn will have more importance than class 3, and so on.

For the Type prediction, there are 3 classes (Attractive, Hotel, and Restaurant). For this reason, we apply the Macro F-measure as the Equation 2 indicates.

$$Res_T(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \quad (2)$$

Where $F_A(k)$ represents the F measure obtained by the system k for the Attractive

class. $F_H(k)$ represents the F measure obtained by the system k for the Hotel class. In the same way, $F_R(k)$ represents the F measure obtained by the system k for the Restaurant class.

Also, for the evaluation of the new subtask, the country classification, the idea is similar to the type prediction measure. The Equation 3 shows the country classification measure.

$$Res_C(k) = \frac{F_{Mex}(k) + F_{Cub}(k) + F_{Col}(k)}{3} \tag{3}$$

The final measure for this task is the average of 3 sub-tasks. The idea is that polarity has more weight than the other two subtasks, it will be given twice the importance, as we can see in the Equation 4.

$$Sent(k) = \frac{2 * Res_P(k) + Res_T(k) + Res_C(k)}{4} \tag{4}$$

To evaluate each system in the unsupervised classification task,that is, the clustering task, an alignment must first be done. Given the Gold Standard, the output of each k system must be renumbered so that the themes correspond. This is because the only restriction that the participating teams have is that they must make 4 groups with the news shared in the competition.

This means that the labels do not necessarily coincide for the same groups expected in the Gold Standard. For this reason, a relabeling will be done for each system using the Gold Standard label that shares the most instances with each of the groups resulting from the k system.

Once the alignment is done, it will be evaluated with a macro F-measure as shown in the Equation 5

$$Thematic(k) = \frac{1}{|L|} \sum_{i=1}^{|L|} F_i(k) \tag{5}$$

Where k is a forum participant system, $L = \{1, 2, 3, 4\}$ and $F_i(k)$ is the F-measure value for the class i obtained by the system k.

2.4 Measuring the easiness of the corpus

As a contribution of this work, we propose a method for measuring corpus easiness. This provides a descriptive and qualitative understanding of the constructed collections.

This measure is based on set operations and is an extension of the measure proposed in (Álvarez-Carmona et al., 2018a). In that work, the difficulty of a paraphrase detection corpus with only two classes is measured. The authors propose counting the words shared by texts that are paraphrased and texts that are not and using that to determine the difficulty of the corpus.

For our problem, we propose something similar. We define a corpus as easier if the texts belonging to one class do not share words with texts from other classes.

Thus, a class is considered easier if it has more exclusive words in its texts. If we define a class as the set of words from the texts that belong to that class, then the easiness of that class, denoted as x, is defined by Equation 6.

$$easiness_{Class}(x, C) = C_x - \bigcup_{i=1, i \neq x}^{|C|} C_i \tag{6}$$

Here, C is the set of classes in the collection, and x is the class for which we need to measure the exclusive elements.

To measure the easiness of the entire corpus, we measure the exclusive words of all classes and divide it by the total number of words in the collection. This is represented by Equation 7.

$$easiness(C) = \frac{\sum_{i=1}^{|C|} easiness_{Class}(i, C)}{\bigcup_{i=1}^{|C|} C_i} \tag{7}$$

Thus, if all words in all classes are exclusive, the value of Equation 7 is 1, indicating the highest easiness index. Conversely, if no class has exclusive words, the value will be 0, indicating the lowest easiness index.

2.4.1 Easiness of the sentiment analysis corpus

Applying the proposed method, as described in Equation 7, yielded the results presented in Table 3. The easiness score for the polarity trait is 0.38, indicating a moderate level of ease. Similarly, the easiness score for the Type trait is 0.50, suggesting a relatively higher level of ease. In the case of the Country trait, the easiness score is 0.47, reflecting a similar level of ease as the Type trait.

Trait	**Class**	$easiness_{Class}$
Polarity	1	5353
	2	5650
	3	13128
	4	31227
	5	81287
	$easiness(Polarity)$	**0.38**
Type	Attractive	72576
	Hotel	52752
	Restaurant	21955
	$easiness(Type)$	**0.50**
Country	Mexico	72044
	Cuba	37765
	Colombia	33751
	$easiness(Country)$	**0.47**

Table 3: Easiness for sentiment analysis.

These results reveal that the type trait is the most straightforward, closely followed by the country trait. Conversely, the polarity trait proves to be the most complex. Among the individual classes, Class 5 demonstrates the highest level of easiness within the Polarity trait, while the Attractive class shows the greatest easiness within the Type trait, and Mexico stands out as the most easily classified within the Country trait.

Conversely, the most challenging classes are Class 1 for Polarity, the Restaurant class within the Type trait, and Colombia within the Country trait.

2.4.2 Easiness of the clustering texts corpus

The same methodology was employed for the Clustering corpus, and the corresponding results are presented in Table 4. The Clustering corpus demonstrated an overall easiness score of 0.61, indicating a potentially higher level of ease compared to the traits in the sentiment analysis task. Notably, the *Insecurity* class emerged as the easiest to classify within the Clustering corpus, while the *Landscapes* class posed the greatest difficulty. However, it is important to note that this task does not have a training phase, therefore, better results will not necessarily be obtained than the traits of the sentiment analysis task.

Group	$easiness_{Class}$
Insecurity	632649
Gastronomy	55891
Prices	37204
Landscapes	25498
$easiness(Clustering)$	**0.61**

Table 4: Easiness for the clustering task.

3 Overview of the Submitted Approaches

This section presents the results obtained by the participants for both tasks.

3.1 Sentiment analysis overview

For this study, 16 teams have submitted 61 solutions for the sentiment analysis task.

Table 5 shows a summary of the results obtained by each team for the sentiment analysis task. This table only shows the best result of each team[3].

The participating teams in the Rest-Mex 2023 sentiment analysis task proposed various approaches to improve classification performance. The LKE-IIMAS team (Murillo et al., 2023) achieved the best results by adapting a RoBERTa Transformer pre-trained with texts in Spanish to the domain of tourism reviews. They also employed data augmentation techniques for minority classes using back translation.

The javier-alonso team (Alonso-Mencía, 2023b) secured second place by using two variants of the RoBERTa model, namely roberta-base-bne and twitter-xlm-roberta-base. They balanced the effect of minority classes through oversampling strategies.

The IIMAS-UNAM team (Baez-Reyes et al., 2023) demonstrated the effectiveness of Transformers with Adapters for sentiment analysis tasks.

The INGEOTEC team (Graff et al., 2023) proposed a multilingual sentiment analysis framework called EvoMSA, which incorporates stacked generalization. They utilized lexical and semantic features to achieve competitive classification results.

The UCT-UA team (Mirabal, Hernández-Alvarado, and Salas, 2023) employed a cascaded approach of transformer-based classifiers biased towards minimizing Mean Average Error for polarity. They also utilized

[3]To see the results of the 61 runs of the forum, you can access `https://sites.google.com/cimat.mx/rest-mex2023/results`

multi-class transformer-based classifiers for predicting type and location.

The BUAA team (Castorena-Salas, Sanchez-Vega, and Lopez-Monroy, 2023) focused on capturing essential features such as writing style, character n-grams, and thematic elements to improve classification performance. They employed the SVM algorithm and conducted a two-stage hyperparameter search.

The Dataverse team (García-Gutiérrez et al., 2023) addressed the challenge of an unbalanced tourist database by using Beto language model-based instances generation for minority classes. They also performed random reduction of instances from majority classes.

The Sena team (Jurado-Buch et al., 2023) constructed a single textual classifier for tourism opinions, incorporating polarity, type, and country of origin. They applied a function to balance the training dataset and employed a Beto transformer for classification.

The UMSNH team (Cerda-Flores et al., 2023) used a late fusion ensemble of multiple methods, including Fasttext, BERT, and micro-TC, adjusted by training subsets and classified using XGBoost.

The JL team (García-Mendoza and Buscaldi, 2023) utilized a BERT-based classification model and data augmentation techniques for minority classes, resulting in improved performance compared to the baseline approach.

The ITT team (Ceballos-Mejá and Álvarez-Carmona, 2023) focused on identifying specific topics related to tourism experiences and employed filtering techniques to classify texts. They used the Beto model for classification.

The Algiedi team (Sandoval, 2023) proposed a reductionist approach, achieving superior classification performance by utilizing a reduced number of instances from the corpus.

The Arandanito team (Carmona-Sánchez, 2023) analyzed the role of verbs, nouns, and adjectives in sentiment analysis of tourist opinions in Spanish.

The LyS-Salsa team (Kellert, Matlis, and Gómez-Rodríguez, 2023) adapted an Unsupervised, Compositional, and Recursive (UCR) rule-based approach to the Universal Dependencies formalism.

The Last team (Gallardo-Hernández and Aranda, 2023) utilized the Mutual Information measure and trained words with normalized MI values as class features.

Two baselines were proposed for this task. The first baseline is the Majority baseline, which simply predicts the majority class for all instances. The second baseline utilizes the Beto model, a variant of BERT, for classification without fine-tuning. By evaluating the performance of Beto without pretraining, the aim is to assess its effectiveness as a standalone model for the task at hand.

It is intriguing to observe that all methods yielded superior results compared to the Majority baseline. Notably, only two teams failed to surpass the Bert-based baseline. This could be attributed to "The Last" team's utilization of a simplistic word count-based method, which may not be as robust as other approaches. Additionally, the LyS-SALSA team only provided predictions for the polarity trait, while submitting random solutions for the other traits. These results signify the valuable findings and high-performing models generated by the participants, showcasing their efficacy in handling tourist text data.

Table 6 presents the top-performing results for each class across the three traits. Notably, the LKE-IIMAS team achieved the highest performance in all classes, except for class 2 of polarity, where the proposal from the javier-alonso team outperformed.

It is intriguing to observe that the minority classes achieved remarkably competitive results, despite the limited number of instances available for training.

These findings suggest a potential correlation between the performance outcomes and the corpus easiness results previously discussed.

3.1.1 Sub-Perfect assemble for the sentiment analysis task

In previous editions of Rest-Mex, as well as in other evaluation forums, the Perfect Ensemble of systems has been computed (Álvarez-Carmona et al., 2018b; Aragón et al., 2019). This approach involves combining all participating systems, where an instance is considered correctly classified if at least one participating system classifies it correctly.

This approach has shown that systems in evaluation forums like this are usually highly

Rank	Institute	Country	Team	Sent	Res_P	Res_T	Res_C
-	-	-	*Perfect Assembly*	0,99	0,99	0,99	0,99
-	-	-	*SubPerfect Assembly*	0,99	0,99	0,99	0,99
1st	BUAP	Mex	LKE-IIMAS$_{RUN_2}$	0,78	0,62	0,99	0,94
2nd	UC3M	Esp	javier-alonso$_{sub_6}$	0,77	0,60	0,99	0,94
3rd	UNAM	Mex	IIMAS-UNAM	0,75	0,59	0,98	0,90
HM	INGEOTEC	Mex	INGEOTEC	0,74	0,55	0,98	0,93
HM	UCT	Chi/Esp	UCT-UA$_{run_1}$	0,72	0,52	0,99	0,94
HM	BUAA	Mex	BUAA$_{M1-M1-M2}$	0,71	0,53	0,98	0,92
HM	CIMAT	Mex	Dataverse	0,71	0,52	0,97	0,92
HM	SENA	Col/Mex	SENA$_{i=1}$	0,70	0,50	0,97	0,88
HM	UMSNH	Mex	UMSNH-xgb$_{ensemble}$	0,69	0,50	0,97	0,89
HM	UC3M	Esp	Camed-CU-ES	0,69	0,51	0,97	0,85
HM	Paris 13	Fra	JL	0,68	0,48	0,97	0,90
HM	ITT	Mex	ITT$_{k=1}$	0,68	0,48	0,96	0,87
HM	Algiedi Solutions	Mex	Algiedi$_{5000}$	0,67	0,49	0,95	0,81
HM	Algiedi Solutions	Mex	Arandanito$_{A+N+V}$	0,63	0,44	0,95	0,78
BL			*Beto-No-Fine-Tuning*	0,38	0,24	0,83	0,34
HM	A Coruña	Esp/Deu	LyS-SALSA	0,31	0,34	0,33	0,32
HM	IBERO	Mex	The-Last	0,23	0,18	0,30	0,25
BL			*Majority*	0,14	0,15	0,20	0,21

Table 5: Performance for the Sentiment Analysis task.

F-measure class	Best result	Team
1	0,68	LKE-IIMAS$_{RUN_1}$
2	0,46	javier-alonso$_{sub_6}$
3	0,58	LKE-IIMAS$_{RUN_2}$
4	0,51	LKE-IIMAS$_{RUN_2}$
5	0,86	LKE-IIMAS$_{RUN_2}$
Attractive	0,99	LKE-IIMAS$_{RUN_2}$
Hotel	0,99	LKE-IIMAS$_{RUN_2}$
Restaurant	0,98	LKE-IIMAS$_{RUN_2}$
Mexico	0,95	LKE-IIMAS$_{RUN_2}$
Cuba	0,94	LKE-IIMAS$_{RUN_2}$
Colombia	0,93	LKE-IIMAS$_{RUN_2}$

Table 6: Performance for the Sentiment Analysis task per class.

complementary and achieve classification effectiveness close to 100%.

However, when combining more than 60 systems, the question arises: Is it possible to select fewer systems that achieve similar performance to the Perfect Ensemble?

To answer this question, we propose a system selection method based on sets. Similar to the method used to measure the easiness of a class (Equation 7), we represent systems as sets, where the instances in each set correspond to the IDs of correctly classified instances.

Thus, a system is considered to contribute more to the ensemble if it has a greater number of exclusively correctly classified elements. This is measured by Equation 8.

$$contributes(k, P) = \frac{P_k - \bigcup_{i=1, i\neq k}^{|P|} P_i}{\bigcup_{i=1}^{|P|} P_i} \quad (8)$$

Using Equation 8, we calculate the extent to which a system exclusively contributes to the ensemble. Systems without unique instances are penalized more than those that contribute more instances, regardless of whether they have many correctly classified elements or not.

The idea is to use this measure to calculate the contribution of each system and eliminate the system that contributes the least to the ensemble.

To evaluate the entire ensemble, we can strive for a balance between the number of correct instances and the number of systems. In other words, an ensemble should be more valuable if it achieves a high result with fewer systems. For this purpose, Equation 9 is proposed.

$$OF(E, \alpha) = |E|^{\alpha-1} \times \sum_{i=1}^{|E|} contributes(k, E) \quad (9)$$

This function gives significant weight to the sum of each contribution in the ensemble and the number of systems. If $\alpha = 1$, the number of systems is not taken into account. If $\alpha > 1$, the equation gives more weight to

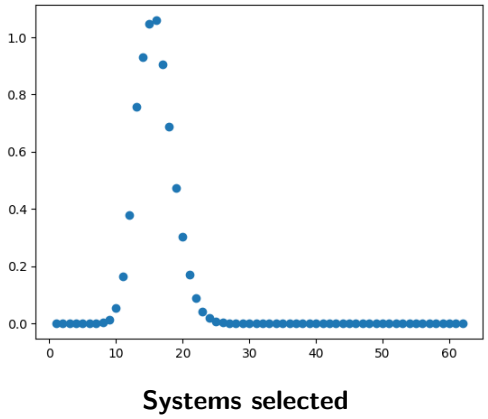

Figure 1: Objective function for different numbers of selected systems.

ensembles with more systems. Conversely, if $\alpha < 1$, the exponent is negative, favoring ensembles with fewer systems. In our case, α should be less than 1.

With this description, Algorithm 1 is proposed.

Data: E, α, increment
Result: E
for i ***in*** $len(E)$ **do**
  s = $[contributes(k, E)$ **for** k **in** $len(E)]$;
  $y.append(OF(s, \alpha + (i \times increment)))$;
  for k ***in*** $len(s)$ **do**
    if $y[k] == \min(y)$ **then**
      $E.pop(k)$;
    end
  end
end

Algorithm 1: Algorithm to select the systems for the assembly.

The *increment* variable is utilized to ensure that as systems are eliminated from the ensemble, the remaining ones gain increased significance or relevance.

The goal of this algorithm is to find the optimal number of systems for the ensemble that maximizes Equation 9.

Figure 1 illustrates the result of Algorithm 1 with $\alpha = 0.5$ and increment $= 1$. As we can see, the optimal number of systems is 16.

Table 5 also presents the results of the Perfect Ensemble (61 systems) and the Sub Perfect Ensemble (16 systems). It can be observed that despite using a significantly smaller number of systems, the results are identical to those of the complete ensemble.

Systems (1-8)	Systems(9-16)
javier-alonso$_{sub2}$	Arandanito$_{V}$
javier-alonso$_{sub8}$	Arandanito$_{N}$
UCT-UA$_{run_1}$	UMSNH-f1weighted$_{ensemble}$
LyS-SALSA	Arandanito$_{A+N+V}$
Arandanito$_{A}$	ITT$_{k=10}$
SENA$_{i=6}$	ITT$_{k=1,1,1,5,5}$
SENA$_{i=4}$	ITT$_{k=-1,-1,2,3,4}$
SENA$_{i=2}$	BUAA$_{M1-M2-M2}$

Table 7: The best 16 systems selected for the assembly.

This demonstrates that the proposed method for system selection performs well. The 16 systems are listed in Table 7.

It is interesting to note that the system that achieved the best results (LKE-IIMAS) is not included in this list. This may be because, although it has the highest number of correctly classified instances, it actually has fewer exclusive instances compared to other systems that achieve lower performance. From this, we can conclude that for a potential late fusion scheme, simply selecting systems with the best results may not necessarily lead to optimal results.

3.1.2 Interesting opinions

Some of the interesting opinions in the collection are those that were correctly classified by all 61 participating systems. These are the easiest instances to classify. The same applies to instances that none of the systems were able to classify correctly, which could indicate the most challenging instances to classify or instances whose labels are incorrect for some reason.

Table 8 shows the number of interesting opinions per class. It can be observed that for polarity, no class was correctly classified by all systems. However, there are instances from classes 2, 3, and 4 that were incorrectly classified by all systems. This is because classifying internal classes can be highly subjective, especially when an instance belongs to class 3 but is actually closer to its neighboring classes like classes 2 and 4. For example:

> **Polarity 3:** *No se pierdan la oportunidad de visitar este hermoso parque, sí les gusta la caminata ó simplemente pasar un buen rato en la naturaleza. este lugar es una de las mejores opciones.*

In the case of the type attribute, it is no-

Class	Correct by All	Wrong by All
1	0	0
2	0	7
3	0	47
4	0	169
5	0	0
Attractive	5320	0
Hotel	0	16
Restaurant	0	12
Mexico	3	0
Cuba	0	28
Colombia	0	16

Table 8: Interesting opinions.

table that the class "Attractive" has 5320 instances that were correctly classified by all systems. This could be because this class was inherently the easiest within this attribute. It is also interesting to observe that there are a few instances from the "Hotel" and "Restaurant" classes that are consistently confused by all systems. Upon examining these instances, it is often found that there are restaurants within hotels, and sometimes tourists mention the restaurant when rating the hotel and vice versa. For example:

> **Hotel:** *Lo mejor es el restaurante. El hotel es bonito pero faltan los detalles de un buen hotel boutique, un trato mas personalizado, pequeños detalles en la habitación, mejores amenidades.*

Finally, there are some opinions about Cuba or Colombia where there is no mention of anything specific to the country, making it challenging for the models to predict the correct destination country. On the other hand, the easiest instances to classify for the label "Mexico" are three instances that discuss the Mayan culture, which appears to be the most distinctive feature compared to Cuba and Colombia. For example:

> **Mexico:** *Pinturas de la historia maya con descripciones en inglés, español y digerir'de'gook. Vale la pena una visita por una hora. Entrada libre.*

3.2 Clustering text results

For this task, 17 systems were received from 4 teams. Table 9 shows the best results of each team[4].

The Javier-Alonso team (Alonso-Mencía, 2023a) proposes a clustering approach that combines K-means and Gaussian Mixture models, leveraging text representations obtained through the multilingual Sentence Transformer. By applying preprocessing techniques such as stopwords and punctuation removal, as well as lemmatization, the team prepares the data. The Sentence Transformer generates text vectors, which are then subjected to dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP). The team identified optimal clusters and fine-tuned hyperparameters by visually analyzing the groups on a two-dimensional UMAP plane. Notably, this approach achieved the best results in the evaluation.

The Cimat team (Rivadenerida-Perez and Callejas-Hernandez, 2023) proposes the utilization of two well-established techniques for conducting thematic modeling, coupled with two robust representation methodologies. Their investigations demonstrate that the most effective segmentation was achieved by combining BERT embeddings and K-means clustering. Additionally, the authors explored the application of LDA and hierarchical clustering using frequency-based representations. Their results secured a commendable second position for this task in the forum. However, it would be intriguing to assess the effectiveness of employing the embedding-based representation in conjunction with the hierarchical clustering approach.

The JCMQ team (Madera-Quintana, Hernández Gónzalez, and Martínez-López, 2023) presents a methodology that utilizes TF-IDF and LSA algorithms to convert texts into vectors. They then employ the K-Means method for text clustering.

The MCE team (Ramos-Zavaleta et al., 2023) employed a two-step approach for clustering the news data. First, they transformed the news using embeddings and subsequently performed dimensional reduction using UMA representation. Then, they applied the OPTICS algorithm as the base method to cluster the data, aiming to enhance accuracy by identifying points that may have been over-

[4]To see the results of the 17 runs of the forum, you can access `https://sites.google.com/cimat.mx/rest-mex2023/results`

Rank	Institute	Country	Team	Thematic	Insec.	Prices	Gastro.	Lscps.
-	-	-	*Perfect Assembly*	0,89	0,97	0,98	0,99	0,62
1st	UC3M	Esp	javier-alonso$_{kmeans10-5}$	0,28	0,61	0,23	**0,21**	0,06
2nd	CIMAT	Mex	CIMAT$_{run3}$	0,24	0,51	**0,24**	0,14	0,06
BL			*Majority*	0,22	**0,91**	0	0	0
3rd	UC	Cub	JCMQ$_{run_5}$	0,21	0,51	0,12	0,06	**0,15**
HM	ITESM	Mex	MCE$_{2ndIter_{Kmeans}}$	0,20	0,50	0,11	0,11	0,07

Table 9: Performance for the Clustering text task.

looked by the base clustering. To further improve clustering results, they stacked a layer of k-means on top of the base cluster results.

As a baseline for this task, only the majority class baseline was considered. This baseline simply predicts the most frequent class, which in this case is "Insecurity." Due to the significant class imbalance in the dataset, the baseline achieves a higher F-measure compared to the JCMQ and MCE teams. However, it should be noted that this is primarily due to the high F-measure of 0.91 obtained in the majority class. Nevertheless, all systems surpass the baseline for the other three classes. The best result for the Prices group was obtained by the Cimat Team. The best result for Gastronomy is obtained by javier-alonso team. Finally, the best result for Landscapes is obtained by JCMQ team.

It is important to emphasize that since this is an unsupervised task, it is not appropriate to evaluate the quality of the clusterings as good or bad. Rather, these four different proposed approaches represent valuable alternatives for various scenarios.

4 Conclusions

This paper described the design and results of the Rest-Mex shared task collocated with IberLef 2023 (Jiménez-Zafra, Rangel, and Montes-y Gómez, 2023). Rest-Mex stands for *Sentiment analysis and clustering Spanish tourists text*. For the two tasks, 20 teams participated. Mainly, the members of these teams come from institutes in countries such as Mexico, Spain, Cuba, Colombia, France, and Chile. 78 systems were received to be evaluated to solve each of the two tasks proposed in the Rest-Mex 2023.

The sentiment analysis task aimed to identify the polarity, type, and country of an opinion. The team that got the best performance was (Murillo et al., 2023). This team represents the University of Puebla in Mexico. They proposed a method based on RoBERTa. This is further evidence of the importance of transformers in textual classification tasks. Also, the results indicate that distinguishing between opinions of hotels, restaurants, and attractions is a task that can have very high results, close to 100 %. The country trait is a little more difficult.

For the clustering task, the best performance is obtained by (Alonso-Mencía, 2023a). The proposed is based on clustering with K-means and Gaussian Mixture. The results show that it is possible to make a grouping on important news related to tourism.

A method was proposed to measure the easiness of a corpus, which appears to effectively reflect the performance of the approaches aiming to solve different tasks.

Finally, it is shown that there is significant complementarity between the participating systems. A system was also proposed to select among the participating systems in such a way that the same results can be achieved by combining only 16 systems, compared to the original 61 systems of the sentiment analysis task.

Acknowledgements

The authors thank the Mexican Academy of Tourism Research (AMIT) for their support of the project "Creation of a labeled database related to tourist destinations for training artificial intelligence models for classifying relevant topics" through the call "I Research Projects 2022", which originated this work.

Our special thanks go to all of Rest-Mex's participants, the organizers, and their institutions.

References

Alonso-Mencía, J. 2023a. Seeking clustering excellence: Unleashing the power of sentence transformers and preprocessing techniques. In *IberLEF@SEPLN*.

Alonso-Mencía, J. 2023b. Unlocking sentiments: Exploring the power of nlp

transformers in review analysis. In *IberLEF@SEPLN*.

Álvarez-Carmona, M. Á., R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, and A. Y. Rodríguez-González. 2021. Overview of rest-mex at iberlef 2021: recommendation system for text mexican tourism. *Procesamiento del Lenguaje Natural*.

Álvarez-Carmona, M. A., R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, and A. D. Pacheco. 2022a. Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University-Computer and Information Sciences*.

Álvarez-Carmona, M. A., R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, and H. Carlos. 2022b. Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news. *Journal of Information Science*, page 01655515221100952.

Álvarez-Carmona, M. Á., Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, and L. Bustio-Martínez. 2022c. Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts. *Procesamiento del Lenguaje Natural*, 69:289–299.

Álvarez-Carmona, M. A., M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, and L. Villaseñor-Pineda. 2018a. Semantically-informed distance and similarity measures for paraphrase plagiarism identification. *Journal of Intelligent & Fuzzy Systems*, 34(5):2983–2990.

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. 2018b. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain*, volume 6.

Aragón, M. E., M. A. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberLEF@SEPLN*, pages 478–494.

Arce-Cardenas, S., D. Fajardo-Delgado, M. Á. Álvarez-Carmona, and J. P. Ramírez-Silva. 2021. A tourist recommendation system: a study case in mexico. In *Mexican International Conference on Artificial Intelligence*, pages 184–195. Springer.

Baez-Reyes, E. Y., I. Barrón-Jiménez, H. Becerril-Pizarro, X. de la Luz Contreras-Mendoza, and I. V. Meza-Ruiz. 2023. Iimas-unam team entry: Transformers adapters for the sentiment analysis rest-mex 2023. In *IberLEF@SEPLN*.

Carmona-Sánchez, N. G. 2023. Measuring the role of the verbs, nouns, and adjectives on the tourist opinions in spanish. In *IberLEF@SEPLN*.

Castorena-Salas, L., F. Sanchez-Vega, and A. P. Lopez-Monroy. 2023. Buaa's team in rest-mex 2023 - sentiment analysis: A basic stylistic and thematic features approach. In *IberLEF@SEPLN*.

Ceballos-Mejá, J. d. J. and M. A. Álvarez-Carmona. 2023. Filtering opinions in spanish with topics of tourist interest for the sentiment analysis task. In *IberLEF@SEPLN*.

Cerda-Flores, J., R. Hernández-Mazariegos, J. Ortiz-Bejar, F. Calderón-Solorio, and J. Ortiz-Bejar. 2023. Umsnh at rest-mex 2023: An xgboost stacking with pretrained word-embeddings over data. In *IberLEF@SEPLN*.

Diaz-Pacheco, A., M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, and R. Aranda. 2022. Artificial intelligence methods to support the research of destination image in tourism. a systematic review. *Journal of Experimental*

& Theoretical Artificial Intelligence, pages 1–31.

Gallardo-Hernández, A. Z. and R. Aranda. 2023. Classifying tourist text reviews by means of mutual information features. In *IberLEF@SEPLN*.

García-Gutiérrez, A. B., P. E. López-Ávila, P. A. Gallegos-Ávila, R. Aranda, and M. A. Álvarez Carmona. 2023. Balancing of tourist opinions for sentiment analysis task. In *IberLEF@SEPLN*.

García-Mendoza, J.-L. and D. Buscaldi. 2023. Enriching with minority instances a corpus of sentiment analysis in spanish. In *IberLEF@SEPLN*.

Graff, M., D. Moctezuma, E. Tellez, and S. Miranda. 2023. Ingeotec at rest-mex: Bag-of-words classifiers. In *IberLEF@SEPLN*.

Guerrero-Rodriguez, R., M. Á. Álvarez-Carmona, R. Aranda, and A. P. López-Monroy. 2021. Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico. *Current issues in tourism*, pages 1–16.

Jiménez-Zafra, S. M., F. Rangel, and M. Montes-y Gómez. 2023. Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org*.

Jurado-Buch, J. D., E. S. Minayo-Díaz, J. A. Tello, K. E. Chaucanes, L. V. Salazar, M. D. Oquendo-Coral, and M. Á. Álvarez-Carmona. 2023. A single model based on beto to classify spanish tourist opinions through the random instances selection. In *IberLEF@SEPLN*.

Kellert, O., N. H. Matlis, and C. Gómez-Rodríguez. 2023. Experimenting with ud adaptation of an unsupervised rule-based approach for sentiment analysis of mexican tourist texts. In *IberLEF@SEPLN*.

Madera-Quintana, J., A. Hernández Gónzalez, and Y. Martínez-López. 2023. Thematic unsupervised classification of tourist texts using latent semantic analysis and kmeans. In *IberLEF@SEPLN*.

Mirabal, P., S. Hernández-Alvarado, and J. A. Salas. 2023. Analyzing sentiment, attraction type, and country in spanish language tripadvisor reviews using language models. In *IberLEF@SEPLN*.

Murillo, V. G. M., H. Gómez-Adorno, D. Pinto, I. A. C. Miranda, and P. Delice. 2023. Lke-iimas team at rest-mex 2023: Sentiment analysis on mexican tourism reviews using transformer-based domain adaptation. In *IberLEF@SEPLN*.

Olmos-Martínez, E., M. Á. Álvarez-Carmona, R. Aranda, and A. Díaz-Pacheco. 2023. What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico. *International Journal of Tourism Cities*.

Ramos-Zavaleta, J., A. Rodriguez, L. Rodriguez, and J. Arreola. 2023. An ensemble based clustering approach to group mexican news. In *IberLEF@SEPLN*.

Rivadenerida-Perez, E. and C. Callejas-Hernandez. 2023. Leveraging lda topic modeling and bert embeddings for thematic unsupervised classification of tourism news in rest-mex competition. In *IberLEF@SEPLN*.

Sandoval, F. 2023. What if we use fewer data to classify tourist opinions in spanish? In *IberLEF@SEPLN*.