Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 289-299 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of Rest-Mex at IberLEF 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts

Resumen de la tarea Rest-Mex en IberLEF 2022: Sistema de Recomendación, Análisis de Sentimiento y Predicción de Semáforo Covid para Textos Turísticos Mexicanos

Miguel Á. Álvarez-Carmona[1,2], **Ángel Díaz-Pacheco**[1], **Ramón Aranda**[1,2], **Ansel Y. Rodríguez-González**[1,2], **Daniel Fajardo-Delgado**[3], **Rafael Guerrero-Rodríguez**[4], **Lázaro Bustio-Martínez**[5]
[1]Centro de Investigación Científica y de Educación Superior de Ensenada
[2]Consejo Nacional de Ciencia y Tecnología
[3]Tecnológico Nacional de México Campus Ciudad Guzmán
[4]Universidad de Guanajuato
[5]Universidad Iberoamericana, Ciudad de México
{malvarez, diazpacheco, aranda, ansel}@cicese.edu.mx
daniel.fd@cdguzman.tecnm.mx, r.guerrero-rodriguez@ugto.mx,
lazaro.bustio@ibero.mx

Abstract: This paper presents the framework and results from the Rest-Mex task at IberLEF 2022. This task considered three tracks: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction, using texts from Mexican touristic places. The Recommendation System task consists in predicting the degree of satisfaction that a tourist may have when recommending a destination of Nayarit, Mexico, based on places visited by the tourists and their opinions. On the other hand, the Sentiment Analysis task predicts the polarity of an opinion issued and the attraction by a tourist who traveled to the most representative places in Mexico. We have built corpora for both tasks considering Spanish opinions from the TripAdvisor website. As a novelty, the Covid Semaphore Prediction task aims to predict the color of the Mexican Semaphore for each state, according to the Covid news in the state, using data from the Mexican Ministry of Health. This paper compares and discusses the participants' results for all three tacks.
Keywords: Rest-Mex 2022, Sentiment Analysis, Covid Prediction, Mexican Tourist Text.

Resumen: Este artículo presenta el marco y los resultados de la tarea Rest-Mex en IberLEF 2022. Esta tarea consideró tres sub tareas: Sistema de recomendación, Análisis de sentimiento y Predicción de semáforo Covid, utilizando textos de lugares turísticos mexicanos. La tarea del Sistema de Recomendación consiste en predecir el grado de satisfacción que puede tener un turista al recomendar un destino de Nayarit, México, con base en los lugares visitados por los turistas y sus opiniones. Por otro lado, la tarea de Análisis de Sentimiento predice la polaridad de una opinión emitida y la atracción por parte de un turista que viajó a los lugares más representativos de México. Hemos construido corpus para ambas tareas teniendo en cuenta las opiniones en español de TripAdvisor. Como novedad, la tarea de Predicción de Semáforo Covid tiene como objetivo predecir el color del Semáforo Mexicano para cada estado, de acuerdo a las noticias Covid en el estado, utilizando datos de la Secretaría de Salud de México. Este documento compara y discute los resultados de los participantes para las tres sub tareas.
Palabras clave: Rest-Mex 2022, Análisis de sentimientos, Predicción de covid, Textos Turísticos Mexicanos.

ISSN 1135-5948. DOI 10.26342/2022-69-26

1 Introduction

Tourism is a social, cultural, and economic phenomenon related to people's movement to places outside their usual place of residence for personal or business/professional reasons (Guerrero-Rodriguez et al., 2021). This activity is vital in various countries, including Mexico (Álvarez-Carmona et al., 2022)[1], where tourism represents 8.7% of the national GDP, generating around 4.5 million direct jobs (Arce-Cardenas et al., 2021).

In 2021, Rest-Mex emerged, which is an evaluation forum (Álvarez-Carmona et al., 2021). This forum is the first that seeks to specialize in text analysis from tourism to provide solutions to different tasks for Mexican Spanish. In its 2021 edition, the Rest-Mex proposed two different tasks. Analysis of recommendation systems and sentiment analysis. For both tasks, data was collected from the TripAdvisor site.

For this Rest-Mex edition, we proposed three sub-tasks: Recommendation System, Sentiment Analysis on Mexican tourist texts, and as a novelty, the task of Determining the color of the Mexican Covid-19 epidemiological semaphore is added.

For this purpose, 3 data sets have been built. We collected **2,263** instances from 2,011 users who visited 18 touristic places in Nayarit, Mexico, for the recommendation system task. For the sentiment analysis task, the data is labeled to determine the polarity and origin of each opinion. For this, **43,150** opinions were collected from various tourist spots in Mexico. Finally, for determining the epidemiologic semaphore, **131,471** news items referring to covid were collected for all the states of the Mexican Republic, grouped into **2,656** weeks.

The remainder of this paper is organized as follows: Section 2 describes this forum's collection-building process and the evaluation metrics. Section 3 summarizes the solutions submitted for the tasks and shows the results obtained by the participants' systems and the analysis. Finally, Section 4 presents the conclusions obtained by this evaluation forum.

[1]Mexico is in the world's top ten and the second Iberoamerican country related to the arrival of international tourists.

2 Evaluation framework

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

2.1 Recommendation System corpus

The first subtask consists of a classification task where the participating system can predict the degree of satisfaction a tourist may have when recommending a destination.

The collection consists of **2,263 instances** with 2,011 tourists and 18 touristic places from Nayarit, Mexico. This collection was obtained from the tourists who shared their satisfaction on TripAdvisor between 2010 and 2020. Each class of satisfaction is an integer between [1, 5], where {1: Very bad, 2: Bad, 3: Neutral, 4: Good,5: Very good}. Each instance consists of two parts:

1. **User information**:
   - Gender: The tourist's gender.
   - Place: The tourist place that the tourist recommends a visit.
   - Location: The place of origin of the tourist (the central, northeast, northwest, west, and southeast regions refer to the regions of Mexico).
   - Date: Date when the recommendation was issued.
   - Type: Type of trip that the tourist would do. The type would be in [Family, Friends, Alone, Couple, Business]
   - History: The history of the places the tourist has visited and his/her opinions on each of these places.
2. **Place information**: A brief text description of the place and a series of representative characteristics of the place as a type of tourism that can be done there (adventure, beach, relaxation, among others.), If it is a family atmosphere, private or public, it is free or paid, among others.

We use a 70/30 partition to divide into train and test. This means that we used

Class	Train instances	Test instances
1	45	20
2	53	24
3	167	72
4	457	196
5	860	369
Σ	1582	681

Table 1: Instances distribution for the recommendation system task.

1,582 labeled instances for the training partition while we used 681 unlabeled instances for the test partition.

Table 1 shows the distribution of the instances for the recommendation system task for the train and test partitions.

The class imbalance is clear since class 5 represents around 50 % of the total instances, making this task very difficult.

Formally the problem of this task is defined as:

"Given a TripAdvisor tourist and a Mexican tourist place, the goal is to automatically obtain the degree of satisfaction (between 1 and 5) the tourist may have when visiting that place."

2.2 Sentiment Analysis corpus

The second subtask is a classification task where the participating system can predict the polarity and the tourist attraction of an opinion issued by a tourist who traveled to the representative Mexican places. This collection was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2021. 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.

The corpus consists of **43,150 opinions** shared by tourists. Like the recommendation task, we use a 70/30 partition to divide into train and test. This means that we used 30,212 labeled instances for the train partition, while we used 12,938 unlabeled instances for the test partition.

Table 2 shows the distribution of the instances for the sentiment analysis task for the train and test partitions for polarity and attraction.

As with the recommendation system subtask, the class imbalance is clear since class 5

Pol			Attr		
Class	Train	Test	Class	Train	Test
1	547	256	Attractive	5197	2216
2	730	315	Hotel	16565	7100
3	2121	884	Restaurant	8450	3622
4	5878	2423	-		
5	20936	9060	-		
Σ	30212	12938	-	30212	12938

Table 2: Instances distribution for polarity and attraction traits on sentiment analysis task.

and the class Hotel represents around 50 % of the total instances, 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 Mexican tourist place, the goal is to determine the polarity, between 1 and 5, of the text and the visited attraction, which could be an attraction, a hotel, or a restaurant."

2.3 Covid Semaphore Prediction

The last subtask is a classification task where the participating system can predict the future of the covid semaphore through the news. This collection was obtained from news websites that published reports regarding covid from June 2020 to December 2021. For this task, **131,471** news items referring to covid were collected for all the states of the Mexican Republic, grouped into **2,656** weeks. Like the previous tasks, a 70/30 partition was made for training and testing. Therefore, 94,540 news items distributed in 1912 weeks were selected for the training corpus. The test corpus consists of 36,931 news items distributed over 744 weeks.

Each week or instance consists of 4 labels. These labels correspond to the semaphore color of the instance after f weeks in the future. The possible colors to detect are: red, orange, yellow, and green, where red is the color that places the most restrictions on public activities and green is the color that corresponds to the best possible situation. The participants must predict the color of the semaphore for the weeks $f = \{0, 2, 4, 8\}$. For more information regarding the covid semaphore, you can consult (Alvarez-Carmona and Aranda, 2022), (Álvarez-Carmona et al., 2022b).

Table 3 shows the distribution of the instances for eaxh f value.

	$f = 0$		$f = 2$		$f = 4$		$f = 8$	
Class	Train	Test	Train	Test	Train	Test	Train	Test
Red	248	87	201	71	179	63	139	42
Orange	680	273	680	275	673	261	655	252
Yellow	545	216	554	221	568	227	615	232
Green	439	168	477	177	492	193	503	218
Σ	1912	744	1912	744	1912	744	1912	744

Table 3: Instances distribution for semaphore prediction.

Like the other tasks, for this corpus, it can be seen that the red class is the minority, which could be the most crucial class to predict, so this task has considerable complexity to solve.

Formally the problem of this task is defined as:

"Given the news set for a week f in a state of the Mexican Republic x, each system must return the color of the covid epidemiological semaphore for weeks f, $f+2$, $f+4$, and $f+8$ for the x state."

2.4 Performance measures

Systems are evaluated using standard evaluation metrics, including accuracy and F-measure, but MAE (mean absolute error) will rank the submissions for the recommendation system task. MAE are defined as equation 1.

$$MAE_{S_x} = \frac{1}{n}\sum_{i=1}^{n}|T(i) - S_x(i)| \qquad (1)$$

Where S_x is a participating system x, $T(i)$ is the result of the instance i according to the Ground Truth, and $S_x(i)$ is the output of the participant system x, for instance, i. Finally, n is the number of instances in the collection.

We proposed a measure to evaluate the sentiment analysis task for this edition. This measure is defined as shown in the equation 2.

$$measure_S = \frac{\frac{1}{1+MAE_p} + F_A}{2} \qquad (2)$$

Where F_A is the average among the micro F-measure for each class (hotel, restaurant, and attractive), and MAE evaluates the polarity.

Finally, for evaluating the semaphore task, we proposed a measure that gives more weight to well-ranked coming weeks to obtain a final result. This measure is defined in the equation 3.

$$measure_C = \frac{F_{w_0} + 2 * F_{w_2} + 4 * F_{w_4} + 8 * F_{w_8}}{15} \qquad (3)$$

Finally, it is essential to mention that the chosen baseline is the majority class for the three tasks.

3 Overview of the Submitted Approaches

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

3.1 Recommendation system overview

For this study, three teams have submitted their solutions for the recommendation system task.

The authors of (Callejas-Hernández et al., 2022) noted that using simpler representations (BoW) independent of the language is well suited for the recommendation task. A similar simple approach is also applied in (Morales-Murillo, Pinto-Avendaño, and Rojas López, 2022). Finally, In (Veigas-Ramírez, Martínez-Davies, and Segura-Bedmar, 2022), a Bert representation is proposed.

Table 4 shows a summary of the results obtained by each team for the recommendation system task. The MAE was used to rank participants. The approach of the GPI-CIMAT team (Callejas-Hernández et al., 2022) obtained the best performance. Surprisingly, the simple approach overcomes the Bert-based approach. This would be the result of the small relativity database. Finally, it was expected that the F-measure of the baseline would not have good results; this is evident since all the experiments surpassed the baseline in this metric, although again, the result exceeded the baseline by 0.05.

Also, Table 4 shows the result of the team that obtained the best result in last year's edition. Since this is the only task of this

edition that is identical to that of the previous year, it is possible to make this comparison. It is possible to see that no team could beat the Alumni-MCE 2GEN team of the 2021 edition (Arreola et al., 2021).

Table 5 shows the best F-measure results by class in the recommendation task. These results show that the minority classes (1 and 2) were not well represented, which is why the best result of the 2021 edition is shown. From class 3, it is possible to see that a different team obtains a good result. It is possible to observe that the GPI CIMAT team obtains the best result for class 5, which explains its better MAE result.

Something remarkable is that all the systems exceeded the baseline (BL).

3.1.1 Perfect assemble for the recommendation system task

To analyze the complementarity of the predictions by the participants' systems, we built a theoretically perfect ensemble (PA) from their runs, as calculated in (Aragón et al., 2019). We considered that a test instance was correctly classified if at least one of the participating teams classified it correctly. Also, it is proposed to combine the participating systems to create a representation based on the outputs of each system. For this, they implemented a deep learning (DL) architecture like the one proposed in (Álvarez-Carmona et al., 2022a).

From these results, it is possible to observe that the perfect ensemble performance is considerably better than the participants' approaches, suggesting that the participants' systems complement each other. This phenomenon has already appeared in this type of task and is known as The Phenomenon of Completeness over Mexican Text Classification (Álvarez-Carmona et al., 2022a).

Figure 1 shows the number of instances correctly classified by s systems. That is, when $s = 0$, all the instances that were not classified well by any system are shown, while when $s = 9$, they are the instances that were classified well by all systems. The base 2 logarithm was applied to the number of instances to observe the graph better. The systems of this and the previous edition were taken for this exercise.

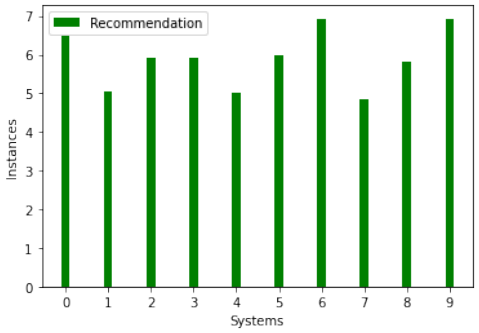

Figure 1: Instances that were correctly classified by the different numbers of possible systems for recommendation.

3.2 Sentiment analysis overview

For this study, 13 teams have submitted their solutions and descriptions for the sentiment analysis task.

For this edition, the transformers-based representation completely dominate the first places for the sentiment analysis task. The UMU team (García-Díaz et al., 2022), UC3M (Pérez Enríquez, Alonso-Mencía, and Segura-Bedmar, 2022), CIMAT MTY GTO (Gómez-Espinosa, Muñiz Sanchez, and López-Monroy, 2022), MCE (Mendoza, Ramos-Zavaleta, and Rodríguez, 2022), GPI CIMAT (Callejas-Hernández et al., 2022), CIMAT 2020 (Santibáñez Cortés et al., 2022), DCI UG (Barco, Rodríguez Rivera, and Hernández-Farías, 2022) and UCI-UC-CUJAE (Toledano-López et al., 2022) implemented solutions, mainly based on Bert. It is possible to observe that these types of methodologies are the ones that obtain the best results since the lowest result, of what transformers applied, is 0.84 when the best result is 0.89, that is, the results are very close to each other.

On the other hand, the rest of the works proposed more straightforward methods. ESCOM-IPN-LCD Team (Alcibar-Zubillaga et al., 2022) proposes a Logistic Regression classifier to train two models, one for polarity prediction and the other for the attraction type prediction. UPTC-UDLAP Team (Rico-Sulayes and Monsalve-Pulido, 2022) applies the Naive Bayes Multinomial algorithm to represent a supervised classification approach. The team uses unigrams, bigrams, and trigrams as features. The unsupervised classification was carried out by computing the total polarity of the opinions in an intensity spectrum according to the scale of the data using context embeddings. The SENA Team (Jurado-Buch, Bustio-Martínez, and Álvarez-Carmona, 2022) uses

Rank	Country	Institute	Team	MAE	Accuracy	F-measure
PA	-	-	-	0.28	86.58	0.66
DL	-	-	-	0.31	76.45	0.52
2021	Mex	CIMAT	Alumni-MCE 2GEN$_{Run1}$	0.31	77.28	0.50
1st	Mex	CIMAT	GPI-CIMAT	**0.69**	52.12	0.19
2nd	Mex	BUAP	LKEBUAP$_{Run_{k1-13-k2-18}}$	0.70	46.64	**0.22**
-	Mex	BUAP	LKEBUAP$_{Run_{k1-14-k2-18}}$	0.70	45.88	0.21
-	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.72	**53.66**	0.17
3rd	Esp	UC3M	UC3M-DEEPNLP$_{Run_1}$	0.72	48.89	**0.22**
BL			Majority Class	0.74	53.30	0.13
-	Esp	UC3M	UC3M-DEEPNLP$_{Run_2}$	0.75	52.71	0.13

Table 4: Performance for the Recommendation System task.

F-measure class	Best result	Team
1	0.32	Alumni-MCE 2GEN$_{Run1}$(2021)
2	0.24	Alumni-MCE 2GEN$_{Run1}$(2021)
3	0.14	UC3M-DEEPNLP$_{Run_1}$
4	0.37	LKEBUAP$_{Run_{k1-13-k2-18}}$
5	0.69	GPI-CIMAT

Table 5: Performance for the Recommendation System task per class.

a representation based on Topics extracted by LDA and classified with simple Deep Learning architecture. Finally, DevsEx-Machina (Rivas-Álvarez et al., 2022) proposes to extract all the terms in each class from one to four words (1...4-grams) as polarity characteristics. Also, they perform a chain of translations of the opinions, from Spanish to other languages and back to Spanish, to obtain meanings and synonymous terms.

It is interesting that despite being more straightforward, some of the results of the proposals that are not based on Transformers obtain close values. This seems ideal for environments with limited memory, time, or data.

Table 6 shows a summary of the results obtained by each team for the sentiment analysis task[2]. The UMU team obtained the best result, although the difference with UC3M is 0.002. Due to the closeness of the results, it is possible that there is no statistical significance between all the methods based on transformers.

Table 7 shows the best F-measure results by class in the sentiment analysis task. Interestingly, the UMU team does not get the best results for any polarity class. However, it is the best team for all three attraction classes. The DCI UG team obtains the best results for classes 1 and 2, which are the most difficult to classify due to their clear imbalance. UC3M obtains the best result for class 3, and finally, MCE, in its two attempts, obtains the best results for the majority of classes.

[2]For systems with *, the authors did not send the system's description.

3.2.1 Perfect assemble for the sentiment analysis task

As in the section 3.1.1, the complementarity of the systems was analyzed for the sentiment analysis task. We calculated the perfect assemble.

Table 6 also shows the perfect assemble (PA) result and the Deep Learning combination systems (DL).

As in the recommendation task, it is possible to observe that the perfect ensemble performance is considerably better than the UMU approach, suggesting that the participants' systems are complementary to each other again, with an error result very close to zero. The DL approach improves the best result obtained by the UMU team.

Figure 2 shows the number of instances correctly classified by s systems similar to the Figure 1. The color green is the polarity instances, whereas the color yellow represents the attraction instances.

3.2.2 Interesting opinions

PA approach got only six incorrect instances for the attractiveness detection task. The pattern of these instances is tourists talking about a hotel restaurant or vice versa, which confuses all systems. For example:

Este lugar era estupendo. Un montón de

Rank	Country	Institute	Team	Measure$_S$	MAE$_P$	F$_A$
PA	-	-	-	0.98	0.03	0.99
DL	-	-	-	0.91	0.19	0.99
1st	Esp	UMU	UMU-Team$_{Run_1}$	**0.89**	**0.25**	**0.99**
2nd	Esp	UC3M	UC3M$_{Run_1}$	**0.89**	0.26	0.98
3rd	Mex	CIMAT	CIMAT MTY-GTO$_{Run_1}$	0.88	0.26	0.98
HM	Mex	ITESM	MCE-Team$_{Run_2}$	0.88	0.26	0.98
-	Mex	ITESM	MCE-Team$_{Run_1}$	0.88	0.26	0.98
-	Esp	UMU	UMU-Team$_{Run_2}$	0.88	0.27	0.98
HM	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.88	0.26	0.98
HM	Mex	CIMAT	CIMAT2020$_{Beto_{Run_1}}$	0.88	0.27	0.97
HM	Mex	INAOE	DCI-UG$_{Run_1}$	0.87	0.26	0.96
HM	Cub/Bel	UCI	UCI-UC-CUJAE$_{Run_2}$	0.87	0.30	0.97
-	Cub/Bel	UCI	UCI-UC-CUJAE$_{Run_1}$	0.86	0.30	0.97
-	Mex	CIMAT	CIMAT2020$_{Run_2}$	0.86	0.31	0.97
-	Mex	INAOE	DCI-UG$_{Run_2}$	0.86	0.30	0.96
HM	Mex	IPN	ESCOM-IPN-IIA*$_{Run_2}$	0.85	0.32	0.96
-	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.84	0.28	0.91
HM	Mex	IPN	ESCOM-IPN-LCD$_{Run_2}$	0.84	0.35	0.94
-	Mex	IPN	ESCOM-IPN-IIA*$_{Run_1}$	0.83	0.34	0.92
HM	Mex/Col	UDLAP	UPTC-UDLAP$_{Run_1}$	0.82	0.44	0.96
HM	Col/Mex	SENA	SENA Team	0.80	0.47	0.92
HM	Mex	UAEM	DevsExMachina$_{Run_2}$	0.70	0.63	0.79
-	Mex	UAEM	DevsExMachina$_{Run_1}$	0.66	0.97	0.82
-	Mex	IPN	ESCOM-IPN-LCD$_{Run_1}$	0.59	0.85	0.65
-	Mex/Col	UDLAP	UPTC-UDLAP$_{Run_2}$	0.54	0.54	0.43
BL	-	-	*Majority class*	0.45	0.47	0.23

Table 6: Performance for the Sentiment Analysis task.

F-measure class	Best result	Team
1	0.61	DCI-UG$_{Run_1}$
2	0.37	DCI-UG$_{Run_1}$
3	0.50	UC3M]
4	0.48	MCE-Team$_{Run_1}$
5	0.88	MCE-Team$_{Run_2}$
Attractive	0.99	UMU-Team$_{Run_1}$
Hotel	0.99	UMU-Team$_{Run_1}$
Restaurant	0.98	UMU-Team$_{Run_1}$

Table 7: Performance for the Sentiment Analysis task per class.

opciones y gran comida fresca. El desayuno buffet era grandes mucha fruta fresca.

This instance is a hotel; however, the opinion refers to the food.

Another example is:

El hotel está increíble, pero resaltó el excelente servicio en insu sky bar, muchas gracias al capitán Iván, y a su staff Heriberto, Gabriel, Luis, Isidoro y Hugo, las bebidas de Gerardo y Alexis increíbles y la cocina un placer ! En el área de alberca al señor Wenceslao! Muchas gracias por todo !

Which, although it is inside a hotel, is a restaurant.

None of these instances have the attractive label.

3.3 Semaphore covid prediction results

For this last task, six systems were received from 4 teams.

MCE team(Ramos-Zavaleta and Rodríguez, 2022) presents an approach based on features extracted directly from the news and the other applying transfer learning. First, they propose a system based on CorEx topics, and as a second attempt, they propose a system based on Bert.

Arandanito team (Carmona-Sánchez, Carmona, and Álvarez-Carmona, 2022) proposes a method based on topic extraction. This topic-based representation is applied to a series of linear regressions, which serve as input for simple deep learning architecture.

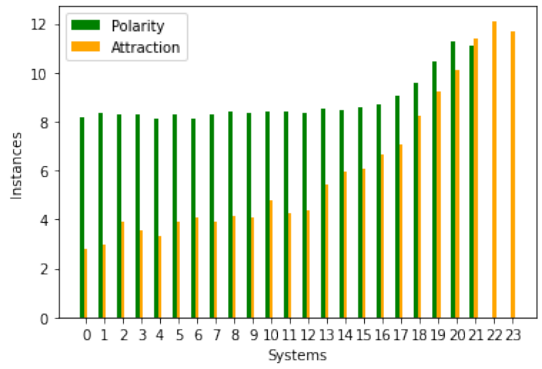

Figure 2: Instances that were correctly classified by the different numbers of possible systems for sentiment analysis.

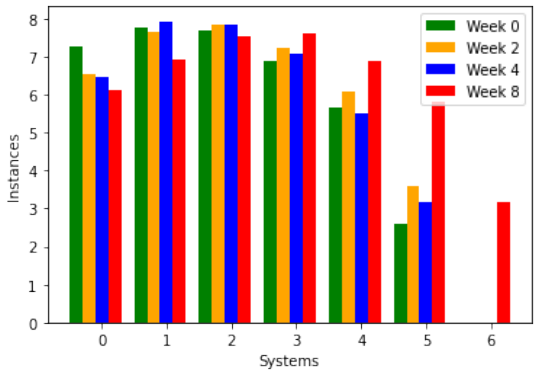

Figure 3: Instances that were correctly classified by the different numbers of possible systems for semaphore prediction.

The last Team (Romero-Cantón et al., 2022) proposes an approach based on weighing representative words as features extracted directly from the text news. Those words were weighed by using the Mutual information (MI) measure.

Table 8 shows the results of the participating teams. It can be seen that both MCE and Arandanito have a very close results to each other. Curiously, both approaches are topic-based.

It can be seen that the best results are obtained for week 2 in the future. That is, two weeks after the news was published. However, the results of weeks 4 and 8, considering the imbalance and that there are four classes, are competitive.

Like the other tasks, all the participants managed to pass the Baseline (BL).

Table 9 shows the best results for each class for each evaluated week.

For week 0, it is possible to see that MCE obtains all the best results. However, from week 2, it can be seen that Arandanito obtains the best result for the Red class. This is the most challenging class because it is the minority class. For all other classes, MCE gets the best result.

3.3.1 Perfect assemble for the semaphore prediction task

Table 8 also shows the perfect assembly and the combination of the systems, like the other two tasks.

The perfect ensemble is much higher than the best of the individual results, which indicates that these systems also complement each other.

On the other hand, the combination of the two systems once again enhances the individual best value.

Figure 3 shows the number of instances correctly classified by s systems similar to the Figure 1. The color green is Week 0, yellow for 2, blue for 4, and red for 8.

For more details of the results of both tasks, it is possible to go to the following web page: `https://sites.google.com/cicese.edu.mx/rest-mex-2022/results`.

4 *Conclusions*

This paper described the design and results of the Rest-Mex shared task collocated with IberLef 2022. Rest-Mex stands for *Recommendation system, Sentiment analysis and covid semaphore prediction in Spanish tourists text for Mexican places*. For the three tasks, 18 teams participated. Mainly, the members of these teams come from institutes in countries such as Mexico, Spain, Cuba, Colombia, Belgium, and Switzerland. Thirty-five different systems were received to be evaluated to solve each of the three tasks proposed in the Rest-Mex 2021.

For the recommendation task, a tourist's satisfaction with a recommendation of a tourist place for the state of Nayarit in Mexico was evaluated. The best MAE result obtained was that of (Callejas-Hernández et al., 2022), which belongs to the CIMAT of Mexico. This team proposed a simple system based on BoW. Although all the participating systems outperformed the Baseline, no system was able to obtain better results than the 2021 edition. This result indicates that this task still has many challenges to be solved.

Rank	Country	Institute	Team	Measure$_C$	F$_{w_0}$	F$_{w_2}$	F$_{w_4}$	F$_{w_8}$
PA	-	-	-	0.84	0.92	0.88	0.87	0.81
DL	-	-	-	0.67	0.74	0.76	0.71	0.63
1st	Mex	ITESM	MCE-Team$_{Run_2}$	**0.49**	**0.56**	0.52	0.46	**0.48**
2nd	Mex	BUAP	Arandanito	0.48	0.33	**0.56**	**0.51**	0.46
-	Mex	ITESM	MCE-Team$_{Run_1}$	0.32	0.33	0.34	0.32	0.32
-	Mex	UNAM	*ML-Team$_{Run_2}$	0.24	0.25	0.27	0.23	0.24
	Mex	UNAM	*ML-Team$_{Run_1}$	0.22	0.20	0.22	0.22	0.23
HM	Mex	UAN	The Last	0.17	0.18	0.18	0.18	0.16
BL			Majority Class	0.12	0.13	0.13	0.12	0.12

Table 8: Performance for the semaphore prediction task.

F-measure class	Best result	Team	F-measure class	Best result	Team
Red$_{w_0}$	0.38	MCE-Team$_{Run_2}$	Red$_{w_2}$	0.37	Arandanito
Orange$_{w_0}$	0.66	MCE-Team$_{Run_2}$	Orange$_{w_2}$	0.68	MCE-Team$_{Run_2}$
Yellow$_{w_0}$	0.45	MCE-Team$_{Run_2}$	Yellow$_{w_2}$	0.52	MCE-Team$_{Run_2}$
Green$_{w_0}$	0.73	MCE-Team$_{Run_2}$	Green$_{w_2}$	0.74	MCE-Team$_{Run_2}$
Red$_{w_4}$	0.39	Arandanito	Red$_{w_8}$	0.2	Arandanito
Orange$_{w_4}$	0.66	MCE-Team$_{Run_2}$	Orange$_{w_8}$	0.65	MCE-Team$_{Run_2}$
Yellow$_{w_4}$	0.47	MCE-Team$_{Run_2}$	Yellow$_{w_8}$	0.55	MCE-Team$_{Run_2}$
Green$_{w_4}$	0.71	MCE-Team$_{Run_2}$	Green$_{w_8}$	0.73	MCE-Team$_{Run_2}$

Table 9: Performance for the Semaphore Prediction task per class.

The sentiment analysis task aimed to identify the polarity and precedence of an opinion made about a Mexican tourist destination. The polarity was evaluated with MAE while the origin with F-Measure. The team that got the best performance was (García-Díaz et al., 2022). This team represents the University of Murcia in Spain. They proposed a method based on Bert. Other teams that also implemented Bert obtained results very close to first place. 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 task of determining the semaphore covid was a novelty introduced for this year's edition. Based on the news regarding covid, this task consists of determining the color of the epidemiological semaphore for weeks 0, 2, 4, and 8 in the future based on the news publications. The best result obtained for this task can be seen in (Ramos-Zavaleta and Rodríguez, 2022). This team comes from ITESM of Mexico. Their solution is based mainly on extracting topics, although they obtains his best result with Bert. Best results are achieved when ranked 2 weeks into the future; however, results for 4 weeks also seem competitive. The results at 8 weeks suffer a drop in the classification.

Finally, it is shown that there is significant complementarity between the participating systems. In other evaluation forums, attempts have been made to mix the participating systems to obtain better results (Álvarez-Carmona et al., 2018), taking the proposal to use a simple deep learning architecture (Álvarez-Carmona et al., 2022a), it was possible to improve the best results of the three tasks. However, the perfect theoretical ensemble is still above the results obtained.

Acknowledgements

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

References

Alcibar-Zubillaga, J., Y. De-Luna Ocampo, I. Pacheco-Castillo, K. Ramirez-Mendez, J. P. M. Sainz-Takata, and O. Juárez Gambino. 2022. Participation of escom's data science group at rest-mex 2022: Sentiment analysis task. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

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. Á., R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, and A. P. López-Monroy. 2022a. A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one. *Computación y Sistemas*, 26(2).

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*.

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. 2018. 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.

Álvarez-Carmona, M. Á., E. Villatoro-Tello, L. Villaseñor-Pineda, and M. Montes-y Gómez. 2022. Classifying the social media author profile through a multimodal representation. In *Intelligent Technologies: Concepts, Applications, and Future Directions*. Springer, pages 57–81.

Alvarez-Carmona, M. Á. and R. Aranda. 2022. Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias.

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.

Arreola, J., L. Garcia, J. Ramos-Zavaleta, and A. Rodríguez. 2021. An embeddings based recommendation system for mexican tourism. submission to the rest-mex shared task at iberlef 2021. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings*.

Barco, G. M., G. E. Rodríguez Rivera, and D.-I. Hernández-Farías. 2022. Sentiment analysis in spanish reviews: Dataket submission on rest-mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Callejas-Hernández, C., E. Rivadeneira-Pérez, F. Sánchez-Vega, A. P. López-Monroy, and E. Villatoro-Tello. 2022. The winning approach for the recommendation systems shared task @rest_mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Carmona-Sánchez, G., A. Carmona, and M. A. Álvarez-Carmona. 2022. Combining linear regressions to determine the future of the covid in mexico from the news. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

García-Díaz, J. A., M. A. Rodríguez-García, F. García-Sánchez, and R. Valencia-García. 2022. Umuteam at rest-mex 2022: Polarity prediction using knowledge integration of linguistic features and sentence embeddings based on transformers. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Gómez-Espinosa, V., V. Muñiz Sanchez, and A. P. López-Monroy. 2022. Automl and ensemble transformers for sentiment analysis in mexican tourism texts. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

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.

Jurado-Buch, J. D., L. Bustio-Martínez, and M. A. Álvarez-Carmona. 2022. The role of the topics for the sentiment analysis task on a mexican tourist collection. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Mendoza, D., J. Ramos-Zavaleta, and A. Rodríguez. 2022. A transfer learning model for polarity in touristic reviews in spanish from tripadvisor. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Morales-Murillo, V. G., D. Pinto-Avendaño, and F. Rojas López. 2022. A hybrid recommender model based on information retrieval for mexican tourism text in rest-mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Pérez Enríquez, M., J. Alonso-Mencía, and I. Segura-Bedmar. 2022. Transformers approach for sentiment analysis: Classification of mexican tourists reviews from tripadvisor. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Ramos-Zavaleta, J. and A. Rodríguez. 2022. A mexico's covid traffic light color prediction system based on mexican news. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Rico-Sulayes, A. and J. Monsalve-Pulido. 2022. A proposal and comparison of supervised and unsupervised classification techniques for sentiment analysis in tourism data. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Rivas-Álvarez, J. C., R. A. García-Hernández, S. I. Medina-Martínez, A. M. Martínez-Ortiz, N. Hernández-Castañeda, J. E. Ruiz-Melo, A. Hernández-Castañeda, and Y. Nikolaevna-Ledeneva. 2022. Devs-ex-machina at rest-mex 2022 opinion mining of the mexican tourism sector through sets of normalized n-grams. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Romero-Cantón, A., A. Diaz-Pacheco, R. Aranda, and P. Ramírez-Silva. 2022. Mexican epidemiological semaphore color prediction by means of mutual information features. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Santibáñez Cortés, E., A. Carrillo-Cabrera, Y. A. Castillo-Castillo, D. A. Moctezuma-Ochoa, and V. H. Muñíz Sánchez. 2022. Bert model and data augmentation for sentiment analysis in tourism reviews for mexican spanish language. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Toledano-López, O. G., J. Madera, H. González, A. Simón-Cuevas, T. Demeester, and E. Mannens. 2022. Fine-tuning mt5-based transformer via cma-es for sentiment analysis. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Veigas-Ramírez, S., D. Martínez-Davies, and I. Segura-Bedmar. 2022. Recommendation system rest-mex 2022 for mexican tourism using natural language processing. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*