Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 163-172 recibido 01-07-2021 revisado 08-07-2021 aceptado 12-07-2021

Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism

Resumen de la tarea Rest-Mex en IberLEF 2021: Sistemas de recomendación para textos turísticos mexicanos

Miguel Á. Álvarez-Carmona[1,2], Ramón Aranda[1,2], Samuel Arce-Cardenas[3], Daniel Fajardo-Delgado[3], Rafael Guerrero-Rodríguez[4], A. Pastor López-Monroy[5], Juan Martínez-Miranda[1,2], Humberto Pérez-Espinosa[1,2], Ansel Y. Rodríguez-González[1,2]
[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]Centro de Investigación en Matemáticas
{malvarez, aranda, ansel, hperez, jmiranda}@cicese.edu.mx
r.guerrero-rodriguez@ugto.mx, pastor.lopez@cimat.mx
{daniel.fd, samuel11290806}@cdguzman.tecnm.mx

Abstract: This paper presents the framework and results from the Rest-Mex track at IberLEF 2021. This track considered two tasks: Recommendation System and Sentiment Analysis, 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 by a tourist who traveled to the most representative places in Guanajuato, Mexico. For both tasks, we have built new corpora considering Spanish opinions from the TripAdvisor website. This paper compares and discusses the results of the participants for both tasks.
Keywords: Rest-Mex 2021, Recommendation System, Sentiment Analysis, Mexican Tourist Text.

Resumen: Este artículo presenta los resultados de la tarea del Rest-Mex en IberLEF 2021. Este evento consideró dos sub tareas, Sistema de Recomendación y Análisis de Sentimientos, ambas utilizando textos turísticos de lugares con interés turístico en México. La tarea del Sistema de Recomendación consiste en predecir el grado de satisfacción que tendrá un turista al recomendar un destino de Nayarit, México, a partir del historial de los lugares visitados por el turista y las opiniones que se le dan a cada uno de ellos. Por otro lado, la tarea de Análisis de Sentimiento consiste en predecir la polaridad de una opinión emitida por un turista que viajó a los lugares más representativos de Guanajuato, México. Para ambas tareas, se han construido dos nuevas colecciones utilizando las opiniones en español del sitio web TripAdvisor. Este artículo compara y analiza los resultados de los participantes para ambas tareas.
Palabras clave: Rest-Mex 2021, Sistemas de recomendación, Análisis de sentimientos, Textos Turísticos Mexicanos.

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 (Di-Bella, 2019). This activity is vital in various countries, including Mexico[1],

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

ISSN 1135-5948. DOI 10.26342/2021-67-14

where tourism represents 8.7% of the national GDP, generating around 4.5 million direct jobs (Elorza, 2020).

With the pandemic generated by the SARS-COV-2 virus, which spread out in Mexico in mid-March 2020, tourism was one of the most affected sectors (Rivas Díaz, Callejas Cárcamo, and Nava Velázquez, 2020). Tourism is trying to re-establish itself through improvements in the quality and safety of touristic products and services (Elorza, 2020).

Natural Language Processing (NLP) is an artificial intelligence area that can help restore tourism by generating mechanisms for detecting problems derived from the identification of polarities in opinions that tourists share on virtual platforms. Systems can also be developed considering the user and destination information to recommend the places where the user may have better tourist experiences. In this way, the tourism sector and the tourists themselves could be supported by the NLP (Anis, Saad, and Aref, 2020).

Few recommendation systems for tourist sites are based on the affinity of a user's profile compared to each place's description. The data collections to train these types of systems are mainly obtained from users and places in English-speaking countries. Considering the relevance of Ibero-American countries for international tourism, it is of utmost importance to generate resources that allow the generation of systems that help develop intelligent systems in Spanish-speaking countries as well.

On the other hand, sentiment analysis tasks in tourist texts has gained relevance in the last decade (Alaei, Becken, and Stantic, 2019). However, as with NLP, the most significant attention of scientific communication efforts have focused on the English language. Although some studies have focused on Spanish, only a few of them address Spanish outside from the country of Spain. These approaches are typically applied to collections taken from social networks such as tweets, so tourist texts have not been directly addressed.

For this Rest-Mex edition, we proposed two sub-tasks: Recommendation System and Sentiment Analysis on Mexican tourist texts.

For this purpose, two 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. As for the sentiment analysis task, 7,413 opinions were collected from tourists who visited Guanajuato, Mexico.

To the best of our knowledge, this is the first time that an evaluation forum is dedicated to solving Tourism issues in Mexican destinations.

The remainder of this paper is organized as follows: Section 2 describes the collection building process for this forum and the metrics for the evaluation. Section 3 summarizes the solutions submitted by the participants for both tasks. Section 4 shows the results obtained by the participants' systems and the analysis. Finally, Section 5 presents the conclusions obtained by this evaluation forum.

2 Evaluation framework

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

2.1 Recommendation System corpus

The first subtask consists in a classification task where the participating system can predict the degree of satisfaction that 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.

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.

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

Table 2.1 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, which makes this a task a very difficult one.

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) that 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 of an opinion issued by a tourist who traveled to the most representative places of Guanajuato, Mexico. This collection was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2020. Each opinion's class is an integer between [1, 5], where:

1. Very negative
2. Negative
3. Neutral
4. Positive
5. Very positive

Each tourist has information about his/her nationality and gender. For example:

- "Un callejón donde tienes que besar a tu amante por años de felicidad, en el amor es parte de un mito en esta ciudad especial. El callejón estrecho con escalones no es muy especial en sí mismo. Lo que lo hace especial es toda la historia a su alrededor"
  - Polarity: 5 (Very positive)
  - Nationality: Mexico
  - Gender: Male
- "Este museo de tres pisos se vende como sede de muchas obras de Diego Rivera, sin embargo, después de recorrer todo el museo, y ante la frustración de no encontrar más que dibujos y bocetos, decidí preguntarle a uno de los guardas, aquí me aclaró que las obras de dos pisos completos se encuentran en restauración, y en otra exhibición en Japón. No dejando así al público ni una sola hora de pintura para apreciar."
  - Polarity: 1 (Very negative)
  - Nationality: Nicaragua
  - Gender; Female

The corpus consists of **7,413 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 5,197 labeled instances for the train partition, while we used 2,216 unlabeled instances for the test partition.

Table 2.2 shows the distribution of the instances for the sentiment analysis task for the train and test partitions.

Class	Train instances	Test instances
1	80	35
2	145	63
3	686	295
4	1596	685
5	2690	1138
Σ	5197	2216

Table 2: Instances distribution for the sentiment analysis task.

As with the other subtask, the class imbalance is clear since, again, class 5 represents around 50 % of the total instances, which makes 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."

2.3 Performance measures

Systems are evaluated using standard evaluation metrics, including accuracy and F-measure, but MAE (mean absolute error) will rank the submissions for both subtasks. 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.

3 Overview of the Submitted Approaches

This section presents the results obtained by the participants for the tasks of recommendation system and sentiment analysis.

3.1 Recommendation system overview

For this study, two teams have submitted their solutions for the recommendation system task. From what they explained in their notebook papers, this section summarizes their approaches regarding pre-processing steps, features, and classification algorithms.

- A Recommendation System for Tourism Based on Semantic Representations and Statistical Relational Learning (Morales-González et al., 2021)
  - **Team:** Labsemco-UAEM
  - **Summary:** The team presented a method of text representation different from the methods of lexical co-occurrence in text. This method extracts the linguistic features in the text, specifically the lexical and semantic signals of synonymy-antonymy. They proposed to use the ComplEx model for the recommendation task. The model was modified to predict the target label, considering it as a relationship between a User and a Place.
- An Embeddings Based Recommendation System for Mexican Tourism. Submission to the REST-MEX Shared Task at IberLEF 2021 (Arreola et al., 2021)
  - **Team:** Alumni-MCE 2GEN
  - **Summary:** The team proposes two methods, the first one is based on Doc2vec. The Doc2Vec model was applied to the user and place information of the dataset. The obtained embeddings were matched with the reviews' centroid embeddings through similarity metrics, and these embeddings were assigned to the design matrix. Finally, for the other user variables, a hot encoding was applied to be incorporated in the design matrix to be modeled through a Neural Network with one hidden layer and ordinal encoding to deal with the unbalanced problem of the data. They proposed a system based on distributed representations of texts for the second method, using the BERT approach.

3.2 Sentiment analysis overview

For this study, seven teams have submitted their solutions and descriptions for the sentiment analysis task. From what they explained in their notebook papers, this section summarizes their approaches regarding pre-processing steps, features, and classification algorithms.

- Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor (Vásquez, Gómez-Adorno, and Bel-Enguix, 2021)
  - **Team:** Minería UNAM
  - **Summary:** They apply two Bert-based approaches for classification. The first approach consists of fine-tuning BETO, a Bert-like model pre-trained in Spanish. The second approach focuses on combining Bert embeddings with the feature vectors weighted with TF-IDF.
- Cascade of Biased Two-class Classifiers for Multi-class Sentiment Analysis (Abreu and Mirabal, 2021)
  - **Team:** UCT-UA
  - **Summary:** The team proposes two methods. The results in their primary submission were obtained from the model BETO. The secondary method has a better result for this team. This method consists of a cascade of binary classifiers based again on BETO.
- DCI-UG participation at REST-MEX 2021: A Transfer Learning Approach for Sentiment Analysis in Spanish (Velazquez Medina and Hernandez Farias, 2021)
  - **Team:** DCI-UG
  - **Summary:** The proposed method is based on a modified Spanish BERT-base architecture model. The BERT-Base architecture was modified by removing the last layer of the network. Then, the last two layers of the modified BERT architecture were concatenated to be used as the input to a dense layer with a swish activation function. As a final layer, a dense layer was used with five outputs (one for each class) using softmax as activation function. For their first run, the model was trained with 70 percent of the training data (with data augmentation for classes 1 and 2) and used the remaining 30 percent as a validation set. On the other hand, the second model was trained using the whole training data (again including the additional data) and the InHouseTest as the validation set to prevent the model from over-fitting.
- Naive Features for Sentiment Analysis on Mexican Touristic Opinions Texts (Carmona-Sánchez, Carmona, and Álvarez-Carmona, 2021)
  - **Team:** Arandanito Team
  - **Summary:** The team proposes a simple method based on naive features, which consist of extracting simple measures such as number of words, number of digits, empty words, among others. They test various classifiers and finally propose a weighting scheme to determine the best classification algorithm; for its representation, it was KNN with $k = 7$.
- Semantic Representations of Words and Automatic Keywords Extraction for Sentiment Analysis of Tourism Reviews (Toledo-Acosta et al., 2021)
  - **Team:** Labsemco-UAEM
  - **Summary:** Firstly, the team proposes an unsupervised method for keyword extraction in order to construct a list of prototypical words conveying a sentiment weight. Secondly, They emphasize the match of the scores of prototypical words with the labels of the texts where they appear. An SVM does the classification task applied to vector representations of text entities.
- Techkatl: A Sentiment Analysis Model to Identify the Polarity of Mexican's Tourism Opinions (Roldan Reyes, 2021)
  - **Team:** Techkatl
  - **Summary:** For this system, the model development and experiments were carried out on the RapidMiner platform. The author proposes filtered stemming words as pre-processing. Their representation is based on TF-IDF. Also, the author applies several classification algorithms. Bayesian Methods obtain the best result.

- Sentiment Classification for Mexican Tourist Reviews based on KNN and Jaccard Distance (Romero-Cantón and Aranda, 2021)
  - **Team:** The last
  - **Summary:** The proposal of this team consists of calculating the Jaccard distance between each instance in the test participation with the average of each of the 5 classes in the train. Jaccard's distance is weighted by the number of repetitions of each word in each class. Finally, the KNN algorithm is used to determine the class of each instance in the test.

4 Experimental evaluation and analysis of results

This section summarizes the results obtained by the participants, comparing and analyzing in detail the performance of their submitted solutions. For the final phase of the challenge, participants sent their predictions for the test partition, the performance on this data was used to rank them. The MAE was used as the primary evaluation measure. In the following, we report the results obtained by participants. Due to the nature of the data being unbalanced, a system that always results in the majority class would have an acceptable result; however, it would not be helpful. For this reason, as a baseline, is proposed the system that always results in class 5 for both tasks.

4.1 Recommendation system results

Table 4.1 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 Alumni-MCE 2GEN$_{Run2}$ team obtained the best performance for all metrics. It is remarkable to observe how this system improved the baseline with 0.42 in MAE. It can also be seen that it surpassed the baseline at 23.37 for accuracy. 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 obtained by the Alumni-MCE 2GEN$_{Run2}$ team exceeded the baseline by 0.37.

Team	MAE	F-measure	Accuracy
Alumni-MCE 2GEN$_{Run1}$	**0.31**	**0.50**	**77.28**
Alumni-MCE 2GEN$_{Run2}$	0.32	0.47	76.21
Baseline	0.73	0.13	53.81
Labsemco-UAEM	1.65	0.16	20.91

Table 3: Performance for the Recommendation System task.

F-measure class	Best result	Team
1	0.32	Alumni-MCE 2GEN$_{Run1}$
2	0.24	Alumni-MCE 2GEN$_{Run1}$
3	0.30	Alumni-MCE 2GEN$_{Run1}$
4	0.67	Alumni-MCE 2GEN$_{Run1}$
5	0.96	Alumni-MCE 2GEN$_{Run1}$

Table 4: Performance for the Recommendation System task per class.

Table 4.1 shows the best F-measure results by class in the recommendation task. In this task, for all classes, the best result was obtained by the same team that obtained the best MAE result, that is, the Alumni-MCE 2GEN$_{Run1}$ team. Unlike which can be intuited, the worst performance class was Class 2 with 0.24, followed by Class 3 with 0.32, when the minority class is the Class 1, which obtained a performance of 0.32. For Class 4, a result of 0.67 was obtained, and finally, for Class 5, which is the majority class, a result of 0.96 was obtained. It should be noted that the baseline of the majority class gets zero from F-measure for all classes, except for Class 5, where it gets 0.69.

4.1.1 Perfect assemble for the recommendation system task

To analyze the complementarity of the predictions by the participants' systems, we built a theoretically perfect ensemble from their runs, as calculated in (Aragón et al., 2019). That is, we considered that a test instance was correctly classified if at least one of the participating teams classified it correctly.

Additionally, we considered a vote approach; we chose the class with the greatest number of predictions among the runs.

Finally, it is essential to mention that 108 instances were not classified correctly by any system. Within these instances, none belong to class 5. On the other hand, 88 instances were correctly classified by all systems. All these instances belong to class 5.

Table 4.1.1 shows the perfect assemble result compared with the best result obtained by the Alumni-MCE 2GEN$_{Run1}$ team. Also, this table shows the vote approach result.

Team	MAE	F-measure	Accuracy
Perfect assemble	0.28	0.77	84.01
Alumni-MCE 2GEN$_{Run1}$	0.31	0.50	77.28
Vote	0.41	0.43	71.84
Baseline	0.73	0.13	53.81

Table 5: Performance for the Recommendation System task.

Team	MAE	F-measure	Accuracy
Minería UNAM$_{Run1}$	**0.47**	0.42	**56.72**
UCT-UA$_{Run2}$	0.54	**0.45**	53.24
UCT-UA$_{Run1}$	0.56	0.40	53.83
DCI-UG$_{Run1}$	0.56	0.28	53.33
Minería UNAM$_{Run2}$	0.58	0.24	54.78
DCI-UG$_{Run1}$	0.60	0.25	53.70
Labsemco-UAEM$_{Run1}$	0.64	0.30	49.05
Techkatl$_{Run1}$	0.66	0.27	50.18
Baseline	0.72	0.13	51.35
Arandanito Team	0.76	0.16	45.71
TextMin-UCLV*$_{Run1}$	0.78	0.17	36.23
Techkatl$_{Run2}$	0.81	0.21	44.76
Labsemco-UAEM$_{Run2}$	0.91	0.24	36.50
TextMin-UCLV*$_{Run2}$	1.00	0.18	38.31
The last	1.26	0.21	36.95

Table 6: Performance for the Sentiment Analysis task.

From these results, it is possible to observe that the perfect ensemble performance is considerably better than the Alumni-MCE 2GEN$_{Run1}$ approach, suggesting that the participants' systems are complementary to each other. Nevertheless, the result from the vote approach indicates that the intersection of correctly classified instances by the systems is relatively small, and therefore, automatically taking advantage of this complementarity is a complex task.

4.2 Sentiment analysis results

Table 4.2 shows a summary of the results obtained by each team for the sentiment analysis task[2]. In total, eight teams with 14 different systems participated. For this task, the Minería UNAM$_{Run1}$ team obtained the best MAE result and the best accuracy; however, the UCT-UA$_{Run2}$ team obtained the best result for F-measure. In this task, eight systems improved the baseline with the MAE measure, 7 improved it in accuracy, and in the same way, as in the recommendation task, all the systems improved the majority class in F-measure.

Table 4.2 shows the best F-measure results by class in the sentiment analysis task. Unlike the recommendation task, different teams obtained the best result for some of the classes. For minority classes like 1 and 2, the best result was obtained by the UCT-UA$_{Run2}$ team with 0.37 and 0.39, respectively. The best results for classes 3 and 4 were obtained by the Minería UNAM$_{Run1}$ team with 0.47 and 0.44, respectively. Finally, the best result for class 5, which is the majority class, was obtained by the Minería UNAM$_{Run2}$ team.

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

F-measure class	Best result	Team
1	0.37	UCT-UA$_{Run2}$
2	0.39	UCT-UA$_{Run2}$
3	0.47	Minería UNAM$_{Run1}$
4	0.44	Minería UNAM$_{Run1}$
5	0.71	Minería UNAM$_{Run2}$

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

Team	MAE	F-measure	Accuracy
Perfect Assenbly	0.06	0.94	96.84
3 best results	0.47	0.47	57.67
Minería UNAM$_{Run1}$	0.47	0.42	56.72
5 best results	0.49	0.39	57.89
8 best results	0.50	0.33	57.53
UCT-UA$_{Run2}$	0.54	0.45	53.24
Baseline	0.72	0.13	51.35

Table 8: Perfect assemble and some combinations for the Sentiment Analysis task.

4.2.1 Perfect assemble for the sentiment analysis task

As in the section 4.1.1, the complementarity of the systems was analyzed for the sentiment analysis task. We calculated the perfect assemble and the vote approaches.

Since there are more participating systems in this task, it is also possible to experiment with vote approaches but with fewer systems. The simple vote approach considers all systems; however, there are systems with results below the baseline, which could be putting more noise in the vote. For this reason, it is proposed to select the approaches to vote concerning the ranking of the table 4.2. In this way, it is proposed to use only the systems above the baseline, that is, the 8 best results. It is also proposed to use the top 5 of systems and finally the top 3.

Table 4.2.1 shows the perfect assemble result compared with the best results obtained by the Minería UNAM$_{Run1}$ and UCT-UA$_{Run2}$ teams. Also, this table shows the vote approaches results.

As in the recommendation task, it is possible to observe that the perfect ensemble performance is considerably better than

the Minería UNAM$_{Run1}$ approach, suggesting that the participants' systems are complementary to each other again, with an error result very close to zero. Nevertheless, the vote approach indicates that the intersection of correctly classified instances by the systems is also relatively small, and therefore, automatically taking advantage of this complementarity is a complex task.

Interestingly, the fewer teams are taken into account for the vote, the better the combination result. This may be because the best systems are taken, and the lower the number of systems, the noise decreases. However, the trend of results indicates that the vote will obtain the same result as the best of the systems in the best cases, making a vote meaningless. Although the accuracy and F-measure statistics were improved in the 3 best results, the MAE measure could not be improved.

4.2.2 Interesting opinions

Two types of interesting opinions can be observed.

1. Opinions that were classified correctly by all systems.
2. Opinions that were not classified correctly by any system.

For the first type, there were 17 opinions in which all systems correctly predicted their class. The 17 opinions belong to class 5. This means that they are very positive, and the text of the opinion clearly shows it. Examples of these opinions are:

- "*Su arquitectura, sus columnas, todo su interior es hermoso su iluminación además de la gente de Guanajuato que lo hacen un lugar mas para visitar*".
- "*Esta basílica es una maravilla tanto en su exterior como interior. Vale la pena conocerla y admirar todos los detalles que tiene.*"
- "*Llegar de noche a este majestuoso lugar, brinda la oportunidad de contemplar una parte bella de la ciudad.*"

For those of the second type, 70 opinions were found that were not correctly classified by any system. It is important to note that none of these opinions are from Class 5. Examples of these opinions are:

- "*En tu visita pasa por ahí es muy especial que lo visites y te enteres de lo que pasa con los cuerpos en ese lugar, es impresionante.*"
  - Class: 1
  - Average of the systems output: 4.71
- "*A todos, este monumento está precioso pero de día hay que visitarle, de noche abstenerse ya que no hay seguridad pública en el lugar y te pueden asaltar.*"
  - Class: 1
  - Average of the systems output: 4.14
- "*Siendo uno de sus atractivos turísticos más importantes, es una lástima la condición en que se encuentra el museo, sucio, sin guías, poca información, encerrado, un decorado sin sentido.* "
  - Class: 1
  - Average of the systems output: 3.57

In the first example, it is clear that the opinion is positive. However, the class awarded by the same tourist is 1 (the lowest). It is possible that the tourist confused the order of the scale, which makes it very difficult to classify this type of opinion correctly. In the second case, the tourist gives a positive opinion but ends with a negative connotation talking about safety issues. Although the word assault (*asaltar*) gives a negative connotation, the other part of the opinion makes the opinion have a higher value; however, the tourist gave it the lowest class. Finally, in the third example, a negative opinion can be observed, but the systems gave a higher rating, possibly due to the bias of the class imbalance towards the more positive classes.

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-2021/results.

5 Conclusions

This paper described the design and results of the Rest-Mex shared task collocated with IberLef 2021. Rest-Mex stands for *Recommendation system and Sentiment analysis in Spanish tourists text for Mexican places.* Two

tasks were proposed, one targeting recommendation tourist places systems and the other focused on sentiment analysis. Mainly, given a set of opinions in Spanish, the participants had to determine the degree of satisfaction that a tourist may have when visiting a Mexican place as well as the polarity of a tourist opinion. For these tasks, we built the two data sets derived from TripAdvisor. The shared task lasted more than three months and attracted 31 teams from countries such as Mexico, Spain, Cuba, Brazil, Chile, Colombia, and the USA. Out of these teams, 9 sent the results of their systems, and 8 sent their report and description of their systems.

The best MAE result for the recommendation task was obtained by (Arreola et al., 2021), while the best result in the sentiment analysis task was obtained by (Vásquez, Gómez-Adorno, and Bel-Enguix, 2021).

For the two tasks, the best results were obtained through representations based on BERT, which again gives evidence that the future of textual classification is directed to the use and application of this type of architecture.

Finally, it is shown that there is significant complementarity between the participating systems of both; however, it does not seem easy to be able to take advantage of the information that each one of them correctly classifies to unite it and improve individual results. This could be an interesting research direction in the future of these tasks.

Acknowledgements

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

References

Abreu, J. and P. Mirabal. 2021. Cascade of biased two-class classifiers for multiclass sentiment analysis. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Alaei, A. R., S. Becken, and B. Stantic. 2019. Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research*, 58(2):175–191.

Anis, S., S. Saad, and M. Aref. 2020. A survey on sentiment analysis in tourism. *International Journal of Intelligent Computing and Information Sciences*, pages 1–20.

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.

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.*

Carmona-Sánchez, G., A. Carmona, and M. A. Álvarez-Carmona. 2021. Naive features for sentiment analysis on mexican touristic opinions texts. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Di-Bella, M. G. 2019. Introducción al turismo.

Elorza, S. R. 2020. Turismo y sars-cov-2 1 en méxico. perspectivas hacia la nueva normalidad. *Desarrollo, economía y sociedad*, 9(1):93–98.

Morales-González, E., D. Torres-Moreno, A. Ehrlich-Lopez, M. Toledo-Acosta, B. Martnez-Zaldivar, and J. Hermosillo-Valadez. 2021. A recommendation system for tourism based on semantic representations and statistical relational learning. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Rivas Díaz, J. P., R. Callejas Cárcamo, and D. Nava Velázquez. 2020. Perspectivas del turismo en el marco de la pandemia covid-19.

Roldan Reyes, E. 2021. Techkatl: A sentiment analysis model to identify the polarity of mexican's tourism opinions. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Romero-Cantón, A. and R. Aranda. 2021. Sentiment classification for mexican tourist reviews based on k-nn and jaccard distance. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Toledo-Acosta, M., B. Martńez-Zaldivar, A. Ehrlich-López, E. Morales-González, D. Torres-Moreno, and J. Hermosillo-Valadez. 2021. Semantic representations of words and automatic keywords extraction for sentiment analysis of tourism reviews. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Vásquez, J., H. Gómez-Adorno, and G. Bel-Enguix. 2021. Bert-based approach for sentiment analysis of spanish reviews from tripadvisor. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*

Velazquez Medina, G. and D. I. Hernandez Farias. 2021. Dci-ug participation at rest-mex 2021: A transfer learning approach for sentiment analysis in spanish. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.*