ISSN 2007-9737

A Combination of Sentiment Analysis Systems for the Study of Online Travel Reviews: Many Heads are Better than One

Miguel Á. Álvarez-Carmona[1,2*], Ramón Aranda[1,2],
Rafael Guerrero-Rodríguez[3], Ansel Y. Rodríguez-González[1,2],
A. Pastor López-Monroy[4]

[1] Consejo Nacional de Ciencia y Tecnología (CONACYT),
Mexico

[2] Centro de Investigación Científica y de Educación Superior de Ensenada,
Unidad de Transferencia Tecnológica,
Mexico

[3] Universidad de Guanajuato,
Mexico

[4] Centro de Investigación en Matemáticas,
Mexico

malvarez@cicese.edu.mx

Abstract. This study presents an analysis of the Rest-Mex forum task 2021, which is the first international evaluation event using tourism-related (Online Travels Reviews - OTRs) data from Mexico. In that forum, 14 specialized sentiment analysis systems were presented. The main contribution of this research is a method to successfully combine those 14 systems specialized on sentiment analysis systems for OTRs. The outputs of those 14 systems were used to evaluate the proposed combination schemes. The systems were trained and tested with 7,413 OTRs from the city of Guanajuato, Mexico, a well-known cultural destination. All of them were collected from TripAdvisor. We propose three schemes to combine the systems to predict the polarity of OTRs efficiently. The combination based on deep learning improves significantly each of the results obtained in the sentiment analysis systems at the individual level. Also, the results were improved for 4 out of the 5 polarity classes in the collection. To the best of our knowledge, this is the first paper that reports results from the combination of different specialized systems in sentiment analysis for OTRs.

Keywords. Sentiment analysis, OTRs, merge systems, deep learning, Mexican tourism.

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 [13]. This activity is vital in various countries, including Mexico, where tourism represents 8.7% of the national GDP, generating around 4.5 million direct jobs [9, 14].

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 [18]. This situation forced several economic sectors to pause their activities, with tourism being one of the most affected causing a disruption at different levels in activities such as accommodation, food services, transport, commerce, among others [12, 14].

Natural Language Processing (NLP) is an artificial intelligence area that has the potential to help in the recovery process of tourism by

doi: 10.13053/CyS-26-2-4055

generating mechanisms for detecting problems derived from the analysis of data from tourists shared on specialized digital platforms such as the case of Online Travels Reviews (OTRs). In this way, the tourism sector and the tourists themselves could be benefited by the NLP [6].

Sentiment analysis tasks in OTRs have gained relevance in the last decade [2, 10, 16, 17]. A significant goal of sentiment analysis is to classify and analyze the polarities of reviews related to products and experiences such as accommodation, online booking sites, e-commerce, social media, among others [25]. However, as with NLP, the most significant attention of scientific communication efforts have focused on the English language mainly. Although it is true that some studies have been conducted on Spanish language, only a few of them address data outside from the country of Spain.

One of the NLP specialized forums that have arisen due to the need to solve tasks related to the analysis of OTRs in Spanish language was Rest-Mex 2021 [3]. Rest-Mex 2021 edition was an international evaluation forum where one of the main tasks proposed by the organizers was to explore sentiment analysis on OTRs from the TripAdvisor website for tourist attractions in the Mexican city of Guanajuato. Among the places under study were the Museum of the Mummies, the University of Guanajuato, the Juarez Theater, the Basilica, the Hidalgo Market, the Union Square, the Alhóndiga de Granaditas and the Kissing Alley.

For this purpose, the organizers collected 7,413 opinions, where 5,197 of them were for the training phase. During this phase, the organizers released these labeled opinions to the scientific community. The participants of this event had the opportunity to build classification models based on machine learning using these 5,197 instances to train learning algorithms in order to predict the polarity of the OTRs in Spanish for the selected tourist attractions. The 2,216 remaining opinions were selected for the test process. This sub-collection was released unlabeled to the participants. The participants classified the instances with their respective models. Finally, participants sent their results obtained to the organizers to be evaluated.

Seven participating teams proposed fourteen different sentiment analysis systems to solve the task. The organizers analyzed the results for each of the systems.

The exciting thing about these results was that the systems are highly complementary. Theoretically, the combination of these systems reaches around 96.8% effectiveness, 56.7% being the best individual result. However, when taking advantage of the information from these systems, their combined result reaches 57.6%. Clearly this is not a substantial improvement considering that multiple systems come together. Therefore, the organizers themselves considered the possibility of finding fusion strategies of the participating results in order to get closer to the best theoretical result.

In this paper, we explore different ways of merging the different participating systems to improve the individual results of different sentiment analysis systems in OTRs in Guanajuato, Mexico.

The aim of this work is to answer the following research questions:

1. Is it possible to merge the Rest-Mex participants systems to improve the results of sentiment analysis for OTRs?

2. Which systems are most important to merge, and what are their characteristics?

3. What tourist attractions can benefit the most from this merging process at the practical level, and what are their characteristics?

The remainder of this paper is organized as follows: Section 2 describes the different international forums specialized in Mexican data and the phenomenon of completeness. Section 3 summarizes and classifies the participants' systems of the Rest-Mex 2021. Section 4 presents the proposal of this work, the methodology, the corpus, the performance measures, and the baselines. Section 5 describe the results obtained, and research questions are answered. Finally, Section 6 presents the conclusions and proposes directions for the future work derived from this study.

2 The Phenomenon of Completeness over Mexican Text Classification Tracks

For many years, NLP leading research focused on the English language considering the scientific community's available data. In order to generate data collections in Spanish language, some organizations such as CLEF, IberEval, or IberLef have organized campaigns to generate evaluation tasks with two main purposes [8,21]: (i) generating data in Spanish for different tasks and (ii) that the scientific community may propose specialized solutions for the Spanish language that can take advantage of this data.

Some of these tasks were proposed exclusively for Mexican Spanish. Among the most important is the Mex-A3t, which proposed to solve the task of author profiling and aggressiveness detection [7], FakeDeS [8] where the fake news detection task is performed, and more recently Rest-Mex [3] being the first to propose a task exclusively to analyze tourism-related data.

The Mex-A3t forum was the first study where the efficacy of the collective result of all participants was measured [4]. This measure was called theoretically Perfect Assembly. During all its editions, the Mex-A3t has reported that this measure gives a result higher than 95%. However, when trying to merge the participating systems, a result is obtained well below the theoretical perfect assembly where they do not even obtain different results to the best participating system. For the Rest-Mex 2021 edition, the same result was also reached. The organizers then proposed a fusion based on vote: first, taking into account all the systems, the best eight systems, the best five, and finally, the best three. However, the best vote result obtained an error of 0.47, where is the same result of the best individual system too.

These evaluation forums' results show that the different participating systems are highly complementary and that the upper bound is the perfect theoretical assembly. However, they have not been able to implement any method so that the fusion of the results surpasses the best-positioned result of each task. This may be because the voting approaches used in [3,4,7,8] are simple. There are more complex fusion methods based on stacking, which is a supervised learning method that learns from the mistakes and successes of the different systems and makes better decisions [5].

This is why this work proposes to generate a fusion of systems based on stacking in such a way that it surpasses at least the best individual results of the Rest-Mex forum when performs the task of sentiment analysis for OTRs.

3 Rest-Mex 2021 Participant Systems on Sentiment Analysis for OTRs

For this study, we propose to merge the solutions of fourteen systems from seven participating teams. This section summarizes and classifies their approaches.

Three different groups of systems were detected. The Transformers-based systems, which apply this type of deep learning architecture. The Bag of Words (BoW) based systems, and Meta-Features based systems.

Table 1 shows the descriptions and types of the participants systems.

4 Methodology

This section describes the proposed fusion, the Rest-Mex corpus characteristics, the baselines, and finally it shows the performance measures.

4.1 Merge methods proposed

There are three different ways of merging the different systems outputs:

1. Weighted vote,
2. Classic Stacking,
3. Deep learning Stacking.

4.1.1 Weighted Vote

For this non supervised method, each system results are merged giving a different level of importance to each system. This level of importance is awarded according to the ranking obtained in Rest-Mex forum. The weighted vote is defined in the following equation 1:

$$Wvote(i) = Max_c \left(\sum_{x=1}^{s} \frac{1}{Rank(S_x)} S_x(i) \right), \quad (1)$$

where s is the number of different systems in the Rest-Mex forum. i is the instance to classify, $S_x(i)$ is the class c returned by the system S_x in the instance i. $Rank(S_x)$ is the ranking obtained by the system S_x in the Rest Mex forum.

In this way, the class chosen, for some instance i will be the one that obtains the majority vote but giving more importance to the systems that obtained the best result.

4.1.2 Classic Stacking

For this method, we propose the application of a supervised approach. This approach takes the different systems' outputs class to generate an array as follows:.

$$arrayStacking(i) =< S_1(i), S_2(i), ..., S_{s-1}(i), S_s(i) > . \quad (2)$$

In this way, an *arrayStacking* can be generated for each instance within the test collection.

Once all the *arrayStacking* are built, it is possible to generate training models that learn from the errors and successes of each of the dimensions in the collection, where each dimension represents a competing system. In this way, classical classification algorithms would determine which systems to consider to obtain the final class. The algorithms that it is proposed to use are SVM, KNN[1], Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB). It is also proposed apply a 10-cross validation scheme for their evaluation.

4.1.3 Deep Learning Stacking

This approach is very similar to classic stacking since *arrayStacking* is generated in the same way as in the equation 2. However, for this variant, it is proposed to use classification based on deep learning. In particular, a neural network with ten hidden layers to ensure that the best relationship is found between the outputs of the participating systems and the real class of each instance. Table 2 summarizes the principal characteristics of the Deep Learning algorithm proposed.

Just like the previous section, it is also proposed to apply a 10-cross validation scheme for their evaluation.

4.2 Rest-Mex Sentiment Analysis Corpus

For the Rest-Mex, the idea was to analyze Online Travel Reviews issued by tourists who visited the most representative tourist attractions in Guanajuato, Mexico. This collection was obtained from the tourists who shared their opinions on TripAdvisor between 2002 and 2020 [3]. Each opinion's class is an integer between [1, 5].

The corpus consists of **7,413 OTRs** shared by tourists. The organizers 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 2,216 were used as unlabeled instances for the test partition [2].

Table 3 shows the distribution of the instances for the sentiment analysis task for the train and test partitions.

Table 4 shows the different attraction in the collection and their polarity average. Also, an OTR example is included per attraction in the original language for illustration purposes. It is possible to observe that there are places that have better rating by tourists. On the other hand, attractions such as Kissing Alley, Hidalgo Market, or the Museum of Mummies are the worst rated. It is also important to mention that the average is 4.27, which is consistent with the class imbalance since negative polarity appears infrequently in the

[1] with $k \in \{1, 3, 5, 7\}$

[2] The corpus is available and can be requested at https://sites.google.com/cicese.edu.mx/rest-mex-2021/corpus-request

Table 1. Systems description

Team	Type	Description
Minería UNAM	Transformers	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 [23].
UCT-UA	Transformers	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. [1]
DCI-UG	Transformers	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 [24].
Labsemco UAEM	BoW	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. [22]
Techkatl	BoW	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 [19].
Arandanito Team	Meta-Features	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 $k = 7$. [11]
The last	Meta-Features	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. [20]

collection. This makes the worst-rated places also the hardest to get good ranking results.

4.3 Performance Measures

Systems are evaluated using standard evaluation metrics, including Accuracy (equation 3), F-measure (equation 5) and MAE (equation 11). All equations involved to measuring the performance of an S_x system are described as follows:

$$Accuracy(S_x) = 100 * \frac{\sum_{i=1}^{n} correct(S_x(i))}{n}, \quad (3)$$

$$correct(S_x(i)) = \begin{cases} 1 \quad If \quad T(i) = S_x(i), \\ 0 \quad Else \quad , \end{cases} \quad (4)$$

$$F - measure(S_x) = \frac{1}{5}\sum_{c=1}^{5} F(S_x, c), \quad (5)$$

$$F(S_x, c) = 2 * \frac{Precision(S_x, c) * Recall(S_x, c)}{Precision(S_x, c) + Recall(S_x, c)}, \quad (6)$$

Table 2. Characteristics of the applied Deep Learning algorithm

Hidden Layers	10
Neurons per layer	1000
Activation function	Relu
Neurons of the final layer	5
Final layer	Softmax
Loss function	Categorical Cross Entropy
Optimizer	Adam
Epochs	50

$$Precision(S_x, c) = \frac{\sum_{j=1}^{n} p_c(j)}{\sum_{i=1}^{|c|} correct(S_x(i))}, \quad (7)$$

$$p_c(j) = \begin{cases} 1 & If \quad T(j) = c, \\ 0 & Else \quad , \end{cases} \quad (8)$$

$$Recall(S_x, c) = \frac{\sum_{j=1}^{n} r_c(S_x(j))}{\sum_{i=1}^{|c|} correct(S_x(i))}, \quad (9)$$

$$r_c(S_x(j)) = \begin{cases} 1 & If \quad S_x(j) = c, \\ 0 & Else \quad , \end{cases} \quad (10)$$

$$MAE(S_x) = \frac{1}{n}\sum_{i=1}^{n} |T(i) - S_x(i)|, \quad (11)$$

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. C is the classes set and $c \in C$. Finally, n is the number of instances in the collection.

Table 3. OTRs instances distribution for the Rest-Mex corpus

Class	Polarity	Train instances	Test instances
1	Very negative	80	35
2	Negative	145	63
3	Neutral	686	295
4	Positive	1596	685
5	Very positive	2690	1138
Σ		5197	2216

4.4 Baselines

As baselines, we propose to use the vote schemes applied by the Rest-Mex organizers. Also, they proposed the majority class as baseline. Finally, the most crucial baseline for our work is the best-ranked result for the forum., This result is obtained by *Minería Unam* team. It is important to mention that the measure the organizers proposed to rank the systems was MAE. For this reason, we will also use it to rank the final results.

5 Experimental Results

Table 5 shows a summary of the results obtained by each team for the sentiment analysis task [3]. For systems with *B* are the baselines, with *P* are the fusion methods proposed; others are normal participants systems of the forum.

The worst results obtained by the proposed methods are those of classic stacking since the best result is obtained by SVM with 0.49 of MAE, 0.34 of F-measure, and 58.03 of Accuracy. Those results are below the majority of baselines and mainly below the best individual result.

The weighted average obtains a better result than SVM. However, it also is below the best individual result.

Finally, only the proposal based on Deep Learning improves all baselines for all metrics, and it is the closest system to Perfect Assembly. For MAE, it is obtained 0.41, 0.58 for F-measure, and 62.59 for Accuracy.

5.1 Analysis of the Results

In this section, we aim to provide answers to the research questions proposed in this study.

[3]* The authors did not send the system's description to the organizers' forum.

5.1.1 Is it Possible to Merge the Rest-Mex Participants Systems to Improve the Results of Sentiment Analysis for Mexican OTRs?

It is possible to merge the participating systems and obtain a considerably better result. However, it is not a straightforward task; it was necessary to resort to one of the approaches that have had the best results in recent years in artificial intelligence: Deep Learning, since other approaches that are also complex were not able to improve the best individual result obtained.

Table 6 shows the best F-measure results by class. It can be seen that the improvement of the merge has a more significant impact on the minority classes. Since for class 5 (very positive), although there is an improvement, it is smaller than for classes 1 (very negative), 2 (negative), and 3 (neutral). Class 4 (positive) was the only one where there was no improvement, and the result of the Minería UNAM team achieved a better result.

5.1.2 Which Systems are Most Important to Merge, and What are their Characteristics?

Table 5 also shows the Information Gain (IG) of each system for the array stacking representation. It is possible to see that, as might be expected, the best systems also have the highest information gain values. This means that those results are the most valuable for the merger. It is even possible to see an inverse correlation between the information gain and MAE of -0.62, a direct correlation of 0.70 with Accuracy, and a very strong direct correlation of 0.93 with F-measure.

It is also clear to see that the transformer-based systems were the most successful both in their individual result and in contributing to the merge results. This could indicate that using only this type of system could help reduce noise and obtain better results.

However, it was not the case. Table 5 also shows the result obtained by the *Deep Learning T* system, which is the same scheme used by the method based on Deep Learning that obtained the best result but only using the systems based on transformers and failed to obtain a result above the baselines. This indicates that although the systems based on BoW and Meta-Features do not have much impact on the merge, their contribution is also essential, and with them, it is possible to surpass all the individual results and baselines. Therefore, it is concluded that although not all systems are equally important, they all provide valuable information.

5.1.3 What Tourist Attractions can Benefit the Most from this Merging Process at the Practical Level, and What are their Characteristics?

The tourist attractions in Guanajuato that obtained a significant improvement when classified using the proposed merge method are Hidalgo Market, the Kissing Alley and the Museum of the Mummies. In particular, the OTRs that have a negative polarity (class 1 and 2), which shows an improvement from 38% of opinions well classified by the best system up to 60% with the proposed method.

This was expected since, as seen in Table 6, the main improvements occurred within these classes. This coincides with the fact that these attractions are among the worst rated by the travelers on TripAdvisor as Table 4 shows.

The correlation coefficient among the average polarity and improved ranking by place is 0.69, which means that while the more negative the overall rating given by tourists to a tourist attraction, the proposed system will have better results than the best individual result.

This supposes clear practical advantages for destinations since the opinions and themes surrounding tourist attractions with negative evaluations from tourists can be detected more efficiently and solutions can be designed accordingly in a shorter period of time. These can range from decision-making by tourism service providers to public policies involving all tourism stakeholders. For more details over themes and problems identified in the collection and these particular tourist attractions in Guanajuato [15].

Table 4. Example instances corpus and polarity average per attraction

Attraction	Polarity Avg	Example
Juarez Theater	4.70	Tomar un tour por este lugar es muy impresionante y bello, la arquitectura, los acabados y toda la historia de lugar es muy interesante
University of Guanajuato	4.60	Me gustó bastante este lugar, solo la pude contemplar desde afuera y me gusto mucho. Definitivamente un Must Visit en Guanajuato.
Union Square	4.59	Hay lugares donde comer muy rico!, de noche es muy bonito y romantico, escuchas a las estudiantinas tocar
Básilica de Guadalupe	4.50	Esta venerable iglesia, ahora basílica de Nuestra Seõra de Guanajuato, es uno de los mejores ejemplos de arquitectura barroca del siglo XVII. En su interior, uno puede admirar la antigua figura de la virgen, Patrona de Guanajuato.
Alhóndiga de Granaditas	4.45	Tiene una variadísima colección de piezas que abarcan siglos, desde antes de la conquista hasta nuestros días. Vale la pena dedicarle tiempo ma apreciar sus murales y las diferentes salas de exhibición.
Pipila Monument	4.27	Nos aconsejaron no ir caminando, es peligroso por los robos a turistas en el trayecto. Se recomienda en visitas guíadas o vehículos
Diego Rivera Museum	4.24	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.
Kissing Alley	3.95	Es un lugar público y como tal se congrega mucha gente solo para tener la oportunidad de tomarse una foto en el famoso callejón, lamentablemente es un caudal de gente que es casi imposible.
Hidalgo Market	3.94	Si vas corto de tiempo, no te molestes en incluir este punto en tu recorrido. Lo que encuentras son playeras con el nombre de la cuidad, artesanías de barro y dulces (que igual los encuentras por toda la ciudad y en cada paseo).
Museum of the Mummies	3.60	Grotesco espectáculo después de dos horas de cola! Vean mejor la película del Santo contra las momias.
Average	**4.27**	

6 Conclusions and Future Work

This work proposed three schemes to merge different systems specialized in sentiment analysis for OTRs. The weighted vote and classic stacking approaches failed to improve the proposed baselines. However, the proposed method based on deep learning surpasses the individual results and the other merge methods.

There is evidence that it is possible to take advantage of the collective information to improve each system. However, given the results, it is

Table 5. Performance of all systems in Sentiment Analysis for OTRs

IG	System	MAE	F-measure	Accuracy	Type
-	*Perfect Assembly*	*0.06*	*0.94*	*96.84*	-
-	Deep Learning*(P)*	**0.41**	**0.58**	**62.59**	-
-	3 best results*(B)*	0.47	0.47	57.67	-
0.307	Minería UNAM$_{Run1}$*(B)*	0.47	0.42	56.72	Transformers
-	Deep Learning T(P)	0.48	0.41	58.43	-
-	Weighted vote*(P)*	0.49	0.39	58.03	-
-	5 best results*(B)*	0.49	0.39	57.89	-
-	SVM*(P)*	0.49	0.34	58.03	-
-	8 best results*(B)*	0.50	0.33	57.53	-
-	KNN-7*(P)*	0.52	0.37	56.00	-
-	NB*(P)*	0.53	0.39	55.68	-
-	RF*(P)*	0.53	0.38	53.70	-
0.299	UCT-UA$_{Run2}$	0.54	0.45	53.24	Transformers
-	KNN-5*(P)*	0.54	0.38	54.10	-
0.272	UCT-UA$_{Run1}$	0.56	0.40	53.83	Transformers
0.149	DCI-UG$_{Run1}$	0.56	0.28	53.33	Transformers
0.125	Minería UNAM$_{Run2}$	0.58	0.24	54.78	Transformers
0.130	DCI-UG$_{Run1}$	0.60	0.25	53.70	Transformers
-	KNN-3*(P)*	0.61	0.34	50.27	-
-	DT*(P)*	0.63	0.33	47.42	-
0.143	Labsemco-UAEM$_{Run1}$	0.64	0.30	49.05	BoW
-	KNN-1*(P)*	0.65	0.29	46.25	-
0.061	Techkatl$_{Run1}$	0.66	0.27	50.18	BoW
-	Majority class*(B)*	0.72	0.13	51.35	-
0.006	Arandanito Team	0.76	0.16	45.71	Meta-Features
0.002	TextMin-UCLV*$_{Run1}$	0.78	0.17	36.23	-
0.008	Techkatl$_{Run2}$	0.81	0.21	44.76	BoW
0.028	Labsemco-UAEM$_{Run2}$	0.91	0.24	36.50	BoW
0.002	TextMin-UCLV*$_{Run2}$	1.00	0.18	38.31	-
0.091	The last	1.26	0.21	36.95	Meta-Features
Correlation	-	-0.62	0.93	0.70	-
Type	Avg IG	Avg MAE	Avg F	Avg Acc	-
Transformers	0.213	0.55	0.34	54.26	-
BoW	0.060	0.75	0.25	45.12	-
Meta-Features	0.048	1.01	0.18	41.33	-

Table 6. Performance per class

F-measure class	Team	Team result	Deep learning result	Improvement
1	UCT-UA$_{Run2}$	0.37	**0.72**	48.62%
2	UCT-UA$_{Run2}$	0.39	**0.54**	27.78%
3	Minería UNAM$_{Run1}$	0.47	**0.51**	7.84%
4	Minería UNAM$_{Run1}$	**0.44**	0.37	-15.9%
5	Minería UNAM$_{Run2}$	0.71	**0.76**	6.57%

concluded that it is not a trivial task and that there is still a wide margin for improvement since the perfect assembly obtains an error result of 0.06 while the approach proposed in this work obtained 0.41.

Although the transformer-based approaches are the most valuable to the mix, all the others also provide valuable information; it is recommended to use all available systems to face this type of task. Collective intelligence is clearly better; in other words ”many heads are better than one”.

The classes that have a significant improvement are the minority. In the sentiment analysis task, the minority classes are usually those with negative polarity (1 and 2), so indirectly, the classification of tourist places that have the most negative OTRs in the collection benefited the most.

The system proposed in this work was applied exclusively to the domain of OTRs related to tourist attractions. However, we consider that it is possible to apply it to multiple types of tourist domains, where OTRs are fundamental such as accommodation and food experiences, service satisfaction, travel and purchase decision-making, destination management, evaluation of destination image and so on. As stated previously in this paper, this type of solution can help to gain a better understanding of the traveling experience directly from the voice of the travelers themselves.

Future work can be directed towards the design and implementation of more complex deep learning architectures so as to get even closer to the result of the perfect assembly. It is considered also essential to explore multilingual collections since these types of mixes are independent of language.

References

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

2. **Alaei, A. R., Becken, S., Stantic, B. (2019).** Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research, Vol. 58, No. 2, pp. 175–191.

3. **Álvarez-Carmona, M. Á., Aranda, R., Arce-Cárdenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A. P., Martínez-Miranda, J., Pérez-Espinosa, H., Rodríguez-González, A. (2021).** Overview of Rest-Mex at IberLEF 2021: Recommendation system for text Mexican tourism. Procesamiento del Lenguaje Natural, Vol. 67.

4. **Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A. (2018).** Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain, volume 6.

5. **Álvarez Carmona, M. Á., Villatoro Tello, E., Montes y Gómez, M., Villaseñor-Pineda, L. (2020).** Author profiling in social media with multimodal information. Computación y Sistemas, Vol. 24, No. 3, pp. 1289–1304.

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

7. **Aragón, M. E., Álvarez-Carmona, M. A., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Moctezuma, D. (2019).** Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. IberLEF@ SEPLN, pp. 478–494.

8. **Aragón, M. E., Jarquín-Vásquez, H. J., Montes-Y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J. P., Bel-Enguix, G. (2020).** Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. IberLEF@ SEPLN, pp. 222–235.

9. **Arce-Cardenas, S., Fajardo-Delgado, D., Álvarez-Carmona, M. Á., Ramírez-Silva, J. P. (2021).** A tourist recommendation system: A study case in Mexico. Mexican International Conference on Artificial Intelligence, Springer, pp. 184–195.

10. **Brahimi, B., Touahria, M., Tari, A. (2020).** Improving Arabic sentiment classification using a combined approach. Computación y Sistemas, Vol. 24, No. 4.

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

12. **Crick, J. M., Crick, D. (2020).** Coopetition and covid-19: Collaborative business-to-business marketing strategies in a pandemic crisis. Industrial Marketing Management, Vol. 88, pp. 206–213.

13. **Di-Bella, M. G. (2019).** Introducción al turismo.

14. **Elorza, S. R. (2020).** Turismo y sars-cov-2 en México. Perspectivas hacia la nueva normalidad. Desarrollo, economía y sociedad, Vol. 9, No. 1, pp. 93–98.

15. **Guerrero-Rodriguez, R., Álvarez-Carmona, M. Á., Aranda, R., López-Monroy, A. P. (2021).** Studying online travel reviews related to tourist attractions using NLP methods: The case of Guanajuato, Mexico. Current Issues in Tourism, pp. 1–16.

16. **Maitama, J. Z., Idris, N., Abdi, A., Bimba, A. T. (2021).** Aspect extraction in sentiment analysis based on emotional affect using supervised approach. 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, pp. 372–376.

17. **Masmoudi, A., Hamdi, J., Belguith, L. H. (2021).** Deep learning for sentiment analysis of Tunisian dialect. Computación y Sistemas, Vol. 25, No. 1, pp. 129–148.

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

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

20. **Romero-Cantón, A., Aranda, R. (2021).** Sentiment classification for Mexican tourist reviews based on K-NN and Jaccard distance. Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.

21. **Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., Stein, B. (2015).** Overview of the PAN/CLEF 2015 evaluation lab. International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp. 518–538.

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

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

24. **Velazquez Medina, G., Hernández Farías, D. I. (2021).** DCI-UG participation at Rest-Mex 2021: A transfer learning approach for sentiment analysis in Spanish. Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings.

25. **Yadav, A., Vishwakarma, D. K. (2020).** Sentiment analysis using deep learning architectures: A review. Artificial Intelligence Review, Vol. 53, No. 6, pp. 4335–4385.

Article received on 03/10/2021; accepted on 19/10/2022.
Corresponding author is Miguel Á. Álvarez-Carmona.