

BUAA's team in Rest-Mex 2023 - Sentiment Analysis: A Basic and Efficient Stylistic and Thematic Features Approach

Lisette Guadalupe Castorena-Salas^{1,*}, Fernando Sánchez-Vega^{2,3} and Adrián Pastor López-Monroy²

¹Benemérita Universidad Autónoma de Aguascalientes (BUAA), Av. Universidad 940, 20100, Aguascalientes, México.

²Mathematics Research Center (CIMAT), Jalisco S/N, Valenciana, 36023 Guanajuato, GTO México.

³Consejo Nacional de Ciencia y Tecnología (CONACYT), Av. de los Insurgentes Sur 1582, Benito Juárez, 03940, CDMX, México.

Abstract

The BUAA team's participation in REST-MEX 2023 and focuses on their exploration of essential features that capture writing style, such as character n -grams, as well as thematic elements like word usage and word n -grams. The Rest-Mex competition encompasses various subtasks, including the identification of tourist attraction types, which involves a thematic perspective. The prediction of the review's country of origin includes significant writing style components that can unveil the specific variant of Spanish used. Lastly, polarity prediction hypothesizes a blend of thematic components or commonly used words to express positive or negative opinions about tourist attractions, while also considering the author's writing style, such as the empathic or friendly tone used to describe the services and amenities of each attraction. The objective was to analyze different stylistic features of the texts and determine the most informative set in order to improve the classification in the analysis of Spanish tourist texts using the SVM algorithm. To achieve this, stylistic and thematic attributes were explored and combined using a two-stage hyperparameter search strategy. With this approach, a sentiment track score of 0.72 was achieved, securing the 6th place out of 17 participating teams. It is important to highlight that this result is significant, considering the simplicity of the proposed solution and a method that requires few computational resources.

Keywords

Rest-Mex 2023, Sentiment Analysis, Bag of Words, Char and word n -grams, Stylistic Features

Introduction

In the digital age, travel experiences are easily shared through digital platforms, providing a wealth of valuable information. User-generated texts, such as social media posts, reviews, and blog articles, offer insights into the sentiments and emotions associated with different destinations. This abundance of information presents an opportunity for sentiment analysis to uncover travelers' attitudes, preferences, and perceptions [1, 2, 3].

IberLEF 2023, September 2023, Jaén, Spain


*Corresponding author.

✉ l.g.c.s0240@gmail.com (L. G. Castorena-Salas); fernando.sanchez@cimat.mx (F. Sánchez-Vega);

pastor.lopez@cimat.mx (A. P. López-Monroy)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The tourism industry recognizes the importance of sentiment analysis in tourism texts. By understanding the emotions expressed by tourists, companies, tourism boards, and researchers can acquire invaluable insights into the factors influencing positive or negative sentiments. This understanding enables stakeholders to make informed decisions, personalize offerings, and enhance the overall travel experience.

Rest-Mex is an international competition that focuses on sentiment analysis of tourism texts in Spanish. The 2023 edition of the competition focuses on sentiment expressed in tourism texts written in the Spanish language obtained from the TripAdvisor platform [4]. Spanish, being one of the most widely spoken languages in the world, provides a vast corpus of user-generated content, allowing us to gain a deep understanding of travelers' experiences and emotions in various regions. The purpose of this shared task competition is to motivate research focused on Spain in the fields of sentiment analysis, detection of variants of Spanish, and specific characteristics of the tourism industry, such as the types of attractions.

The Rest-Mex competition problem is to determine the polarity, ranging from 1 to 5, of opinions about tourist attractions. Additionally, we aim to predict the type of tourist place visited, such as a hotel, restaurant, or attraction, as well as the country that was visited, whether it is Mexico, Colombia, or Cuba [4].

Traditionally, sentiment analysis has been conducted by observing the words used in opinions. Specific dictionaries or word sets associated with negative or positive polarity have been employed, as well as ad hoc selections obtained through attribute selection techniques [5, 6]. While the words themselves can indicate polarity, how they are used must also be taken into account. To address this aspect, we propose the incorporation of simple stylistic attributes [7] to discern the author's style when determining the polarity expressed in a review. These simple style attributes have proven effective in identifying polarity within contexts where texts exhibit informality or lack meticulousness, such as social networks (as observed in [8]). Moreover, they have found widespread applicability in languages with limited resources [9, 10], which further underscores their utility. Hence, we advocate leveraging these simple style attributes, (specifically char n -grams) for sentiment analysis in the Spanish language. Furthermore, the integration of these attributes holds significant potential for enhancing the recognition of Spanish language variants since they reflect the stylistic features adopted by authors depending on the variant they use.

We believe that all subtasks could benefit from characterizing the writing style. Additionally, we emphasize the importance of considering thematic aspects and sentiment analysis keywords. To ensure comprehensive coverage of both thematic and stylistic characteristics, we have employed character n -grams for capture these attributes. Furthermore, in order to reinforce the thematic elements and polarity keywords, we have additionally incorporated attributes obtained through a traditional bag-of-words approach [11].

The proposed method, utilizing the concatenation of optimized bags of words and characters, has demonstrated its efficiency and competitive performance when compared to computationally demanding deep learning approaches. This approach achieved an the 6th place ranking among a multitude of deep learning proposals. By combining carefully optimized word and character representations, our method captures a comprehensive range of influences, offering an effective alternative that mitigates the need for extensive computational resources.

1. Task and corpus description

In this section, a detailed description of the tasks carried out to participate in the "Rest Mex 2023" sentiment analysis competition will be presented. Additionally, a comprehensive explanation will be provided regarding the corpora used in the research, which were crucial for developing and evaluating sentiment analysis models.

1.1. Task

Given a specific opinion or review about a tourist destination, the main objective is to determine the polarity or sentiment expressed in the text on a scale ranging from 1 to 5. The numerical scale represents the spectrum of sentiment, where a rating of 1 indicates the highest degree of dissatisfaction, while a rating of 5 reflects the highest degree of satisfaction.

In addition to analyzing the polarity of sentiment, the competition also aimed to predict the type of tourist site visited by the reviewer. The target categories include hotels, restaurants, and attractions, allowing for a more comprehensive understanding of sentiment across different aspects of the tourism industry.

Furthermore, the competition sought to identify the origin of the reviewer, focusing on three specific countries: Mexico, Colombia, and Cuba. By determining the geographical origin of the reviewers, the analysis could potentially reveal cultural nuances and variations in sentiment expression, providing valuable insights into regional perceptions and preferences.

Through the Rest-Mex 2023 competition [4], participants were challenged to develop innovative methodologies and machine learning models that can effectively address these aspects of sentiment analysis in Spanish tourism texts. The ultimate goal was to accurately classify sentiment polarity, determine the type of tourist site, and identify the origin of the reviewer, thus enriching our understanding of travelers' emotions, preferences, and perceptions across various destinations.

1.2. Development corpus: Rest-Mex 2022

The database for the 2022 [12] competition consists of 30,212 instances that were created since 2021 [13]. Some collaborators searched for the 30 most relevant tourist places in Guanajuato and Jalisco, such as hotels, restaurants, and attractions, from the official TripAdvisor website. Each instance contains the following information:

1. Title: the title that the tourist assigned to their review.
2. Opinion: the opinion expressed by the tourist.
3. Polarity: the polarity of the opinion: [1, 2, 3, 4, 5].
4. Attraction: the type of place for which the opinion is being expressed: ["Hotel", "Restaurant" or "Attractive"].

Furthermore, the polarity ranges from 1, indicating the highest degree of dissatisfaction, to 5, representing the highest degree of satisfaction. It can be interpreted as follows: Very bad (1), Bad (2), Neutral (3), Good (4), Very good (5).

1.3. Competence corpus: Rest-Mex 2023

In 2023 [4], the database from 2022 [12] was reused, and in addition, the 30 most relevant tourist places from Puebla, Nuevo León, Veracruz, Cuba, and Colombia were added. This resulted in a training dataset of 251,702 instances. Each instance contains the following information:

1. Title: the title that the tourist assigned to their review.
2. Opinion: the opinion expressed by the tourist.
3. Polarity: the polarity of the opinion: [1, 2, 3, 4, 5].
4. Country: the country that was visited.
5. Type: the type of place for which the opinion is being expressed: ["Hotel", "Restaurant" or "Attractive"].

Just like in the previous corpus, the polarity range is from 1 to 5.

2. Methodology

This section provides a detailed exposition of the methodology used in the investigation. It thoroughly describes the steps and procedures followed, as well as the strategies used to achieve reliable and significant results, and verify the relevance of the stylistic characteristics in the different subtasks, as well as the convenience of joining these with the thematic characteristics of the texts.

Preprocessing

Particularly, our dataset consists of reviews, so we need to preprocess them. The preprocessing steps are described as follows:

- Tokenization: involves dividing a document into words (or character chunks if is the case).
- Lowercasing: The idea is to replace all uppercase letters with lowercase letters.
- Stop words: Common words in the document that do not provide much information can be removed.

Text Representation

To apply machine learning tools, we need to generate a representation of the documents. Here are descriptions of some used representations:

- n -grams: are sequences of n continuous words or characters extracted from the text.
- Use idf: Enabling this option allows for inverse document frequency weighting.
- Smooth idf: It adds one to the document frequency as if there was an additional document containing all terms from the collection only once.
- Sublinear tf: It replaces tf with $1 + \log(tf)$.
- Norm: There are two types of norms: $L2$ norm, which ensures the sum of squared vector elements is 1, and $L1$ norm, which ensures the sum of absolute values of vector elements is 1 [14].

Feature Selection

- Max features: Builds a vocabulary based on the given top frequency.
- Min df: Ignores terms in the document that have a frequency lower than the specified threshold.

The addition (or omission) of each of these techniques and the parameters with which they are used define the set of hyperparameters of the representation. The subsection below describes the hyperparameter selection methodology followed.

Procedure for Hyperparameter selection

In this subsection, we will outline the procedure followed for selecting the hyperparameters in our study. Hyperparameters play a crucial role in determining the performance of machine learning models, and their optimal selection is essential for achieving accurate and robust results. However, exhaustive exploration of all possible combinations through a greedy search becomes computationally infeasible and time-consuming. Therefore, to address this challenge and optimize the parameters effectively, a systematic strategy is essential, enabling an efficient search approach that balances computational resources and performance optimization.

We start by identifying the relevant hyperparameters for our model and their potential range of values. This involved selecting two groups of hyperparameter characteristics and evaluating the model's performance using each combination. The selection of hyperparameters was carried out in two stages. First, the best values for the hyperparameters of group 1 were chosen with group 2 fixed. Then, the values of group 2 were optimized using the best configurations from group 1. This approach was taken because evaluating all hyperparameter combinations would require an impractically large grid, exceeding our available resources. The evaluations were performed using the 2022 [12] database for polarity and attraction type tasks since the number of instances was optimal for testing. However, for country prediction, we used a 20% subset of the 2023 [4] training database since only Mexican texts were available in 2022.

We used appropriate standard evaluation metrics such as $F1$ score (also known as F -measure) and Mean Squared Error (MAE) [15] to assess the model's performance for each hyperparameter combination. To ensure reliable performance evaluation and reduce the risk of overfitting, we applied four-fold cross-validation technique. This technique allowed us to evaluate the performance of different hyperparameter combinations on multiple data subsets and obtain more robust results.

The explored combinations of hyperparameters covered various dimensions. Firstly, we examined different units of analysis, including word-based analysis related to thematic attributes, as well as individual characters and characters surrounded by word boundaries or white spaces, which are associated with stylistic attributes. Additionally, we considered other preprocessing features, such as case inclusion/exclusion, stop word removal, minimum frequency threshold, and the option to apply inverse frequency transformation or not.

Regarding word n -grams, we considered a wide range of options, ranging from unigrams to 5-grams, and also explored combinations between them, e.g., unigrams, bigrams, and trigrams. For character n -grams, we set a minimum range of 2 characters and a maximum of 9, exploring all combinations within that range. Additionally, the maximum feature limit was set at 25,000.

After analyzing the results based on the selected evaluation metric on the second stage, we identified the combinations that yielded the best performance with different ranges of word and character n -grams. Then, we kept the features of these selected models fixed and performed additional combinations by varying the maximum feature.

In addition to variations in the maximum feature limit, we also explored different norms for model processing, considering options such as $L1$, $L2$, or none. Lastly, we made decisions regarding the use of additional text representations, such as smooth IDF and sublinear TF.

The features of the top four models from the second group of hyperparameter combinations were selected. The thematic and stylistic-oriented representation attributes were then merged. Finally, the best final models were obtained by combining the combinations that yielded the best evaluations and were trained using the entire training dataset.

By following this procedure, we aimed to find the optimal set of hyperparameters that maximized our model's performance. This iterative process allowed us to fine-tune the model and improve its effectiveness.

The diagram in Figure 1 summarizes the aforementioned process, illustrating the workflow followed in the hyperparameter selection.

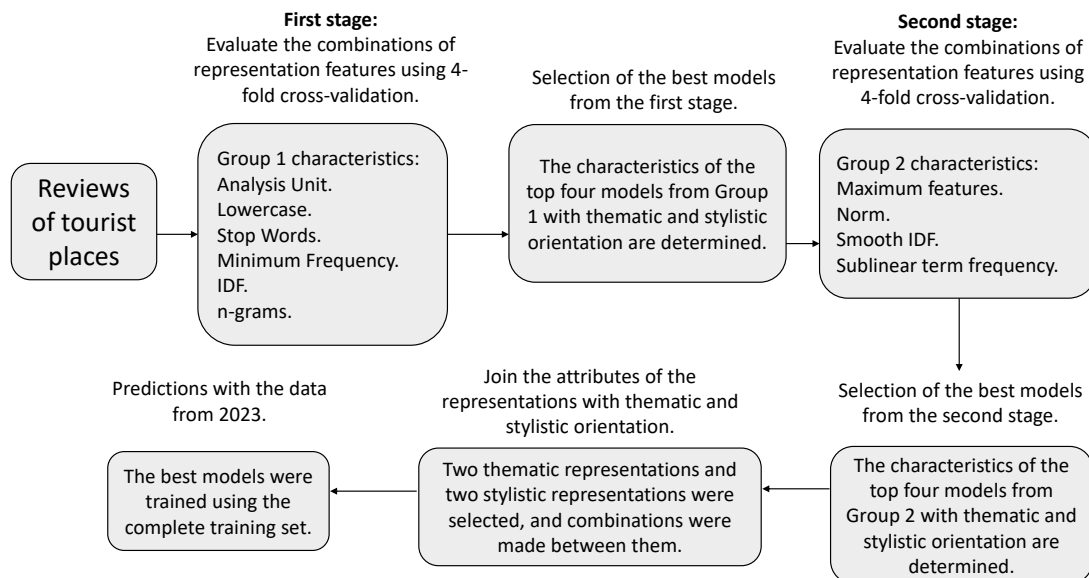


Figure 1: Diagram of the research development. Own elaboration.

Machine learning algorithm

The algorithm applied for classification was Support Vector Machine (SVM) with a linear kernel. The linear kernel allows the model to capture linear relationships between features and sentiments.

3. Results

In this section, we present the results of the experiments conducted in the research on sentiment analysis of tourism texts, as well as the results obtained in the Rest-Mex 2023 competition.

3.1. Preliminary evaluation and hyperparameter setting

Next, we present the main findings and results obtained during the experiments. These results are based on common evaluation metrics such as MAE and F-measure.

First Stage

In the first stage of the experiments, we explored the hyperparameter combinations from Group 1. These combinations involved varying different units of analysis (character, character with word boundaries, and words), inclusion/exclusion of capitalization, removal of stop words, minimum frequency threshold (1, 3 or 5), option to apply inverse frequency transformation or not, and the aforementioned ranges of n -grams. It's worth noting that the maximum feature limit was fixed at 25,000 for this stage. A total of 1584 hyperparameter configurations were generated, and each combination was evaluated based on the F -measure. Figure 2 displays the average, variability, maximum, and minimum values of the F -measure for each hyperparameter while varying the remaining configurations. This analysis helps determine the range of coupling between each hyperparameter and the rest of the hyperparameters in Group 1.

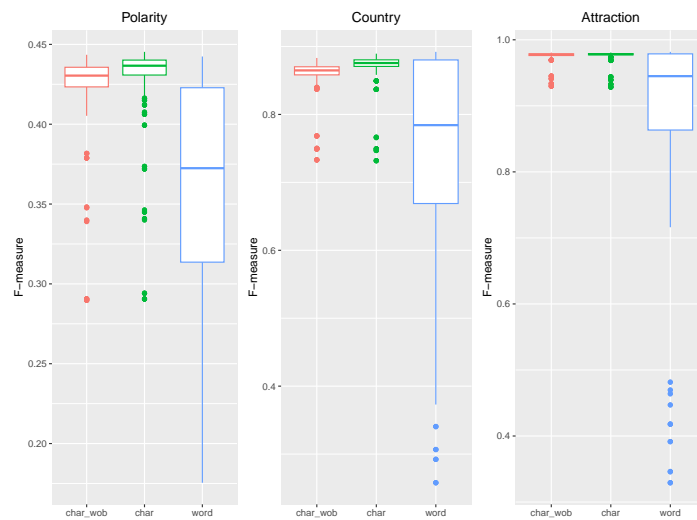


Figure 2: Box plot displaying the task evaluations in the first stage, differentiated by the unit of analysis used. Own elaboration.

Noticeable differences in the average $F1$ scores among the three units of analysis are observed, with the "character" attribute analysis obtaining the highest average. To provide a clearer view of the maximum value, Figure 3 presents a zoomed-in representation.

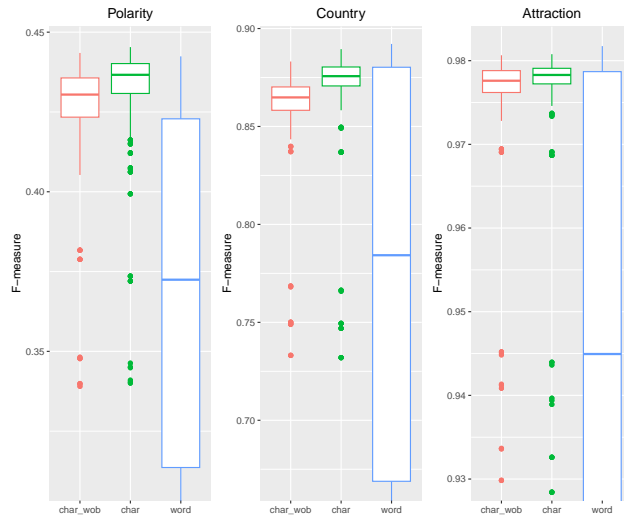


Figure 3: Zoom-in of Figure 2 showing the task evaluations in the first stage, differentiated by the unit of analysis used. Own elaboration.

In Figure 4, some outliers and the minimum value were omitted since the majority of evaluations regarding the use of hyperparameters such as stop words, minimum frequency, lowercase letters, and inverse frequency transformation are concentrated in the upper part of the graph. On average, the best results were obtained when using the IDF weighting scheme and maintaining uppercase letters. Additionally, no significant differences were found when adjusting the minimum frequency and considering stop words.

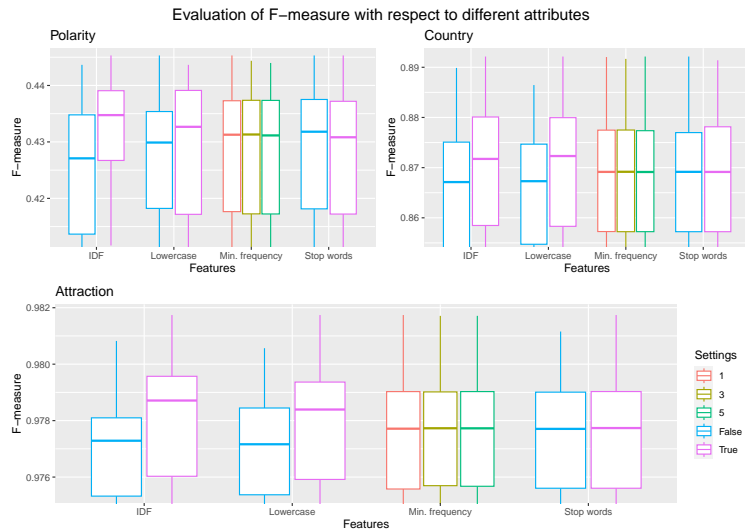


Figure 4: Zoom-in to the evaluation of the algorithm with F -measure regarding different features for the first stage in the three tasks. Own elaboration.

These results highlight the importance of carefully selecting features and settings in sentiment analysis of tourism texts. Additionally, they provide valuable insights into the factors influencing the accuracy and effectiveness of sentiment analysis models applied to this specific domain.

In Figure 5, the most relevant n -gram ranges are presented, with the limits of the graph varying according to the task.

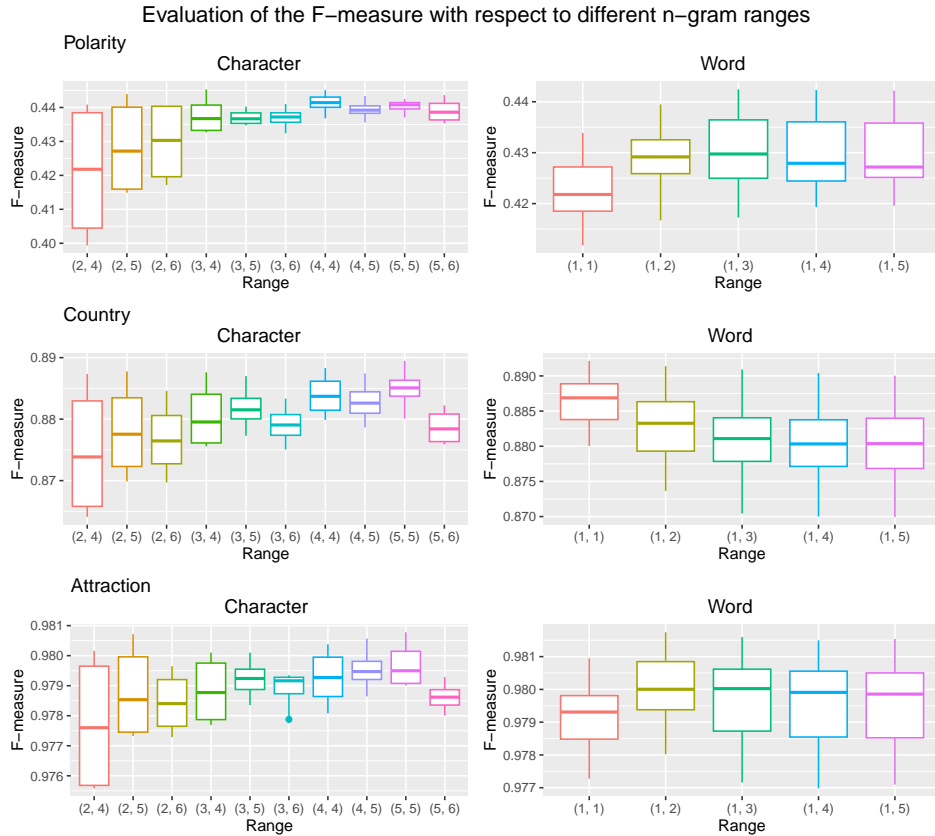


Figure 5: Evaluation of the algorithm in different n -gram ranges. Own elaboration.

For polarity prediction, it is observed that the bag of words containing 3-gram and 4-gram characters, as well as single 4-gram characters, achieve the highest $F1$ scores. In general, the 4-gram characters show the best average. However, for word n -grams, it was found that the n -grams consisting of 1, 2, and 3 words together yield better scores.

Regarding the country and attraction tasks, the results are very similar for character n -grams, with the 5-gram characters having the highest average. However, for word ranges, it was found that the best model for the country task is based on word unigrams, while for the attraction task, a combination of unigrams and bigrams achieves a better score.

Based on these results, selected ranges were chosen, one with a smaller n -gram range and another with a wider range, which showed the best values in terms of the F-measure. These selected ranges, along with the aforementioned attributes, were kept fixed for use in the second stage.

Table 1 presents the hyperparameters for each task that yielded the best results in the first stage of the experiments and were kept fixed in the second stage. The first column represents each task, the second column indicates the unit of analysis, and the remaining columns denote "T" (True) to indicate that the hyperparameter was used, or "F" (False) if it was not.

Table 1
Best configurations of the first stage.

Task	Analyzer	Lower-case	Stop words	<i>n</i> -grams	Min. frequency	Use idf	F1
Polarity	char	F	T	(3, 4)	1	T	0.445317608
	char	F	F	(2, 5)	5	T	0.443948821
	word	T	F	(1, 3)	1	F	0.442421945
Contry	word	T	F	(1, 1)	5	T	0.892136835
	word	T	T	(1, 5)	5	T	0.890028849
	char	T	F	(5, 5)	5	T	0.889476643
	char	T	T	(2, 5)	1	T	0.887768425
Attractive	word	T	T	(1, 2)	1	T	0.981741058
	word	T	T	(1, 5)	3	T	0.981537794
	char	T	F	(5, 5)	5	T	0.980776476
	char	T	F	(2, 5)	5	T	0.980711607

Second stage

In the second stage of the experiments, additional analyses were conducted with the aim of refining and improving the results obtained in the first stage. The following are the most relevant findings and results from this stage.

In the first stage, certain hyperparameters were used to analyze the polarity task, including different units of analysis (characters and words), from which three of the best configurations were selected (the settings shown in the Table 1). In the second stage, additional hyperparameters were incorporated, such as different norms (L1, L2, or none) and variations in the maximum number of features in the bag-of-words (5000, 10000, 25000, 30000, and 60000). Different smoothers, such as smooth IDF and sublinear TF, were also combined. In total, 180 combinations were obtained for the polarity task.

For the prediction of country and attraction, the best hyperparameter configurations were also selected in the first stage, but in this case, two were chosen for character-based analysis and two for word-based analysis. In the second stage, the same hyperparameters as in the polarity task were added, resulting in a total of 240 combinations for these tasks.

Figure 6 and Figure 7 show the variation in the use of Sublinear TF and Smooth IDF smoothers when changing the maximum number of tokens or features in the bag-of-words matrix, respectively.

In general, we observe that as the maximum number of features increases, the *F1* score also tends to increase in both cases of smoothers. This indicates that a larger number of features allows for capturing more relevant information in sentiment analysis of tourism texts.

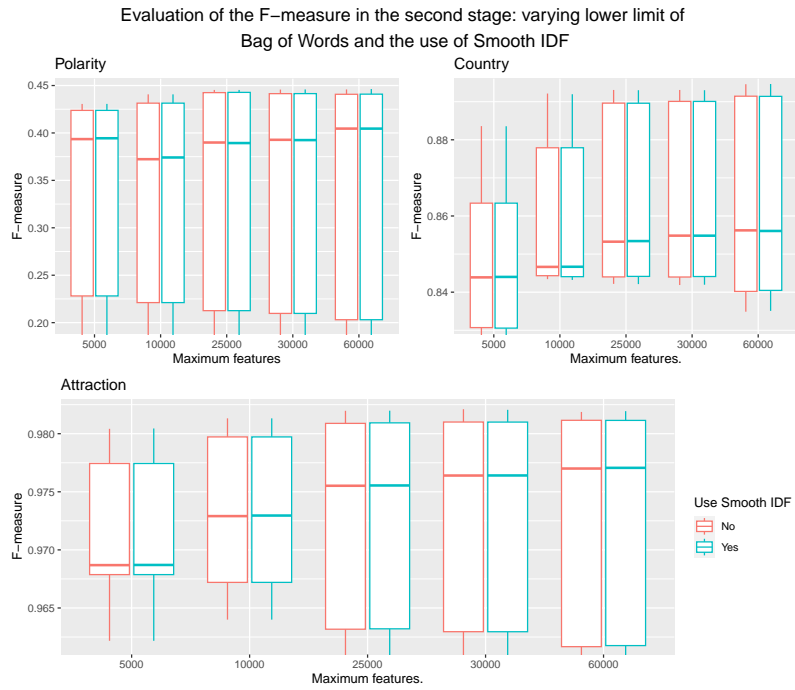


Figure 6: Zoom-in to the box plot of the F-measure varying the maximum number of features for the use of Smooth IDF. Own elaboration.

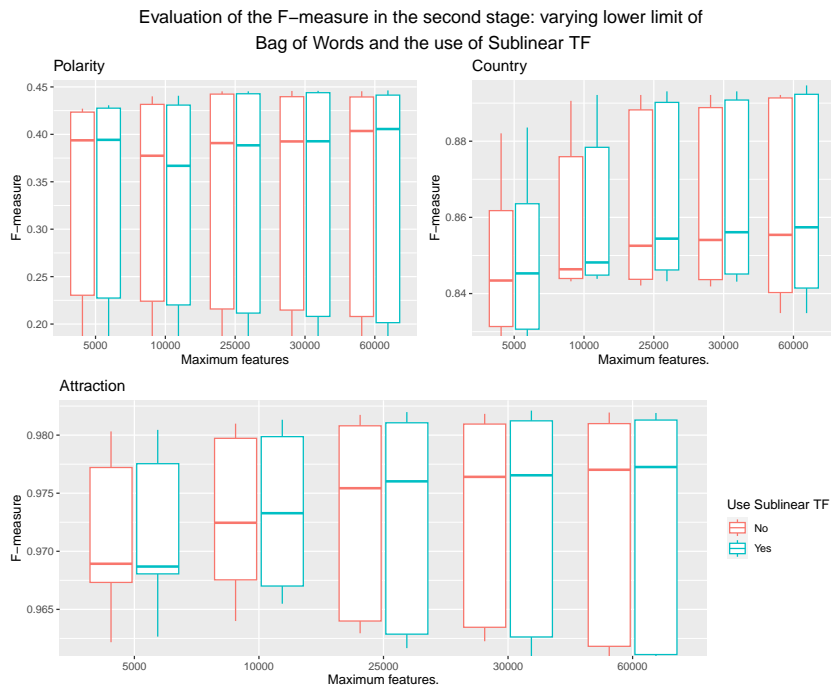


Figure 7: Zoom-in to the box plot of the F-measure varying the maximum number of features for the use of Sublinear TF. Own elaboration.

Table 2 summarizes the best configurations identified for each specific task, considering both the representation features used in the first stage and the additional features incorporated in this stage.

Table 2
Best configurations of the experiments.

Task	ID	Ana-lyzer	Lower-case	Stop words	n-grams	Min. freq.	Use idf	Max. feat.	Norm	Smooth idf	Subli-near tf
Polarity	Pch1	char	F	T	(3, 4)	1	T	60000	l2	T	T
	Pch2	char	F	F	(2, 5)	5	T	60000	l2	F	F
	Pw1	word	T	F	(1, 3)	1	F	60000	l2	T	T
	Pw2	word	T	F	(1, 3)	1	F	25000	l2	F	T
Country	Cch1	char	T	T	(2, 5)	1	T	60000	l2	F	T
	Cw1	word	T	F	(1, 1)	5	T	25000	l2	F	T
	Cch2	char	T	F	(5, 5)	5	T	60000	l2	F	T
	Cw2	word	T	T	(1, 5)	5	T	60000	l2	F	T
Attractive	Aw1	word	T	T	(1, 2)	1	T	30000	l2	F	T
	Aw2	word	T	T	(1, 5)	3	T	60000	l2	F	T
	Ach1	char	T	F	(5, 5)	5	T	60000	l2	F	T
	Ach2	char	T	F	(5, 5)	5	T	30000	l2	F	T

These optimal configurations were selected based on their performance in terms of the F -measure. By presenting these configurations in a table format, it facilitates the comparison and selection of the most effective options for each particular task.

Table 3 shows the F -measure obtained for the best configurations along with their respective ID.

Table 3
 F -measure of the experiments in the second stage

Task	ID	F1
Polarity	Pch1	0.446310430
	Pch2	0.445448316
	Pw1	0.445118245
	Pw2	0.444209386
Country	Cch1	0.894644467
	Cw1	0.893081132
	Cch2	0.892597473
	Cw2	0.892196580
Attractive	Aw1	0.982107636
	Aw2	0.981873273
	Ach1	0.981243284
	Ach2	0.981212856

After selecting the best configurations from the second stage, the attributes of the topic-

oriented and stylistic representations were combined to evaluate if better performance could be achieved in the models. The results obtained for each of the tasks are presented below (Tabla 4).

Table 4
F-measure of the configuration unions

Settings	<i>F</i> 1 - Country (C)	<i>F</i> 1 - Attraction (A)	<i>MAE</i> - Polarity (P)	<i>F</i> 1 - Polarity (P)
[Xch1, Xw1]	0.893735	0.983282	-0.349398	0.458224
[Xch1, Xw2]	0.897709	0.983277	-0.350887	0.458767
[Xch2, Xw1]	0.892514	0.983187	-0.348636	0.457856
[Xch2, Xw2]	0.897624	0.983332	-0.349629	0.457873
[Xw1, Xw2]	0.894561	0.981384	-0.355289	0.446682
[Xch1, Xch2]	0.891168	0.980621	-0.360883	0.447545
[Xw1, Xw2, Xch1, Xch2]	0.894651	0.982923	-0.357375	0.456111
[Xw1, Xw2, Xch1]	0.896173	0.983243	-0.35519	0.453721
[Xw1, Xw2, Xch2]	0.895477	0.98307	-0.353899	0.455634
[Xch1, Xch2, Xw1]	0.891874	0.982855	-0.354197	0.457497
[Xch1, Xch2, Xw2]	0.89623	0.982921	-0.355984	0.455321
Xch1	0.894587	0.981213	-0.357441	0.44631
Xch2	0.892597	0.981154	-0.35562	0.445448
Xw1	0.893081	0.982108	-0.346584	0.445105
Xw2	0.892197	0.981873	-0.349927	0.444205

3.2. Competition Results

Finally, the following combinations were submitted from the best models, and the results obtained are shown in Table 5.

Table 5
Sentiment Track Score Rest-Mex 2023

Polarity	Attractive	Country	Sentiment Track Score
[Pch1, Pw1]	[Ach1, Aw1]	[Cch2, Cw2]	0.718921943
[Pch1, Pw1]	[Ach2, Aw2]	[Cch2, Cw2]	0.718858763
[Pch1, Pw1]	[Ach1, Aw1]	[Cch1, Cw1]	0.718757375
Pw1	Aw1	Cw1	0.713125921
Pch1	Ach1	Cch1	0.708782446
BaseLine-Beto-No-Fine-Tuning			0.381064303
Average			0.623386446

4. Conclusion

The BUAA team’s participation in REST-MEX 2023 focused on exploring essential features that capture writing style and thematic elements to improve the classification of Spanish tourist texts using the SVM algorithm. When comparing the two types of explored attributes, character n-grams and words, we found that character n-grams perform better in predicting the destination

and polarity in terms of the F-measure, while word n-grams are slightly better in predicting the type of tourist place and considering the MAE for polarity. These findings were expected, as speaker recognition relies more on stylistic elements, while identifying the type of tourist attraction is more thematic. On the other hand, for polarity identification of individual classes, character n-grams appear to be more effective, while for modeling polarity trend using MAE, words are a slightly better option. It was also found that combining stylistic and thematic attributes with the union of bag-of-words from both characters and words always allows for better classification with a higher F1 score.

In the sentiment analysis competition, we achieved a score of 0.72, ranking 6th out of 17 teams. This score surpasses the benchmark set by BERT (BaseLine-Beto-No-Fine-Tuning) and is above the average of the participants. Our approach proved to be successful in the sentiment analysis competition for tourist texts, achieving a notable score and outperforming most of the competitors. Also, this result is particularly noteworthy given the simplicity of the proposed solution and the minimal computational resources required.

In conclusion, these results highlight the importance of considering different types of attributes and their combination in sentiment analysis of tourism texts. By incorporating both stylistic and thematic attributes, as well as using bag-of-words from characters and words, the classification accuracy is improved, leading to better results in terms of the F1 measure. These findings provide a solid foundation for the development of more accurate and effective sentiment analysis models in the tourism domain, enabling companies and organizations in the industry to make informed decisions and enhance the travelers' experience.

5. Ethical Issues

We consider it highly relevant to address tasks concerning languages from the global south, which, despite having large populations, often lack attention in the development of language technologies. While we acknowledge the value of exploring linguistic variants and their identification, it is essential to acknowledge that such methodologies can inadvertently perpetuate market segmentation and biases in the provision of tourist and cultural offerings across countries and populations with divergent socioeconomic levels.

Additionally, it is noteworthy that the proposed methodology in this study operates within resource constraints, yet still achieves competitive outcomes when compared to approaches necessitating access to robust computing systems and graphical processing units (GPUs). Hence, it is imperative to maintain research endeavors that prioritize sustainable, straightforward, and efficacious methodologies, thereby enabling widespread adoption among populations with limited access to the requisite hardware resources for deep learning techniques.

Acknowledgments

We sincerely thank the Municipality of Rincón de Romos, and the Government of Aguascalientes, through INCyTEA, for their valuable financial support. Thanks to their generosity, we have been able to carry out our research and present our paper at the IBERLEF 2023 Congress in Jaén, Spain. Their financial backing has been crucial to our academic success, and we look

forward to continuing our collaboration in the future to promote research and development in our community. Sanchez-Vega would like to thank CONACYT for its support through the Program “*Investigadoras e Investigadores por México*” by the project No. 1311, ID. 11989.

References

- [1] R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, *Current Issues in Tourism* (2021) 1–16. doi:<https://doi.org/10.1080/13683500.2021.2007227>.
- [2] M. A. Álvarez Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Ángel Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, *Journal of King Saud University - Computer and Information Sciences* 34 (2022) 10125–10144. URL: <https://www.sciencedirect.com/science/article/pii/S1319157822003615>. doi:<https://doi.org/10.1016/j.jksuci.2022.10.010>.
- [3] A. Diaz-Pacheco, M. A. Álvarez Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, *Journal of Experimental & Theoretical Artificial Intelligence* 0 (2022) 1–31. doi:[10.1080/0952813X.2022.2153276](https://doi.org/10.1080/0952813X.2022.2153276).
- [4] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, *Procesamiento del Lenguaje Natural* 71 (2023).
- [5] M. D. Molina-González, E. Martínez-Cámara, M. T. Martín-Valdivia, S. M. J. Zafra, esolhotel: Generación de un lexicón de opinión en español adaptado al dominio turístico, *Proces. del Leng. Natural* 54 (2015) 21–28. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5090>.
- [6] A. Moreno-Ortiz, C. P. Hernández, Lexicon-based sentiment analysis of twitter messages in spanish, *Proces. del Leng. Natural* 50 (2013) 93–100. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4664>.
- [7] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. Carvalho, E. Stamatatos, Authorship attribution for social media forensics, *IEEE Trans. Inf. Forensics Secur.* 12 (2017) 5–33. URL: <https://doi.org/10.1109/TIFS.2016.2603960>. doi:[10.1109/TIFS.2016.2603960](https://doi.org/10.1109/TIFS.2016.2603960).
- [8] Q. Han, J. Guo, H. Schütze, Codex: Combining an SVM classifier and character n-gram language models for sentiment analysis on twitter text, in: M. T. Diab, T. Baldwin, M. Baroni (Eds.), *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, The Association for Computer Linguistics, 2013, pp. 520–524. URL: <https://aclanthology.org/S13-2086/>.
- [9] I. Habernal, T. Ptáček, J. Steinberger, Sentiment analysis in czech social media using

- supervised machine learning, in: A. Balahur, E. V. der Goot, A. Montoyo (Eds.), Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2013, 14 June 2013, Atlanta, Georgia, USA, The Association for Computer Linguistics, 2013, pp. 65–74. URL: <https://aclanthology.org/W13-1609/>.
- [10] J. Kapociute-Dzikiene, A. Krupavicius, T. Krilavicius, A comparison of approaches for sentiment classification on lithuanian internet comments, in: J. Piskorski, L. Pivovarova, H. Tanev, R. Yangarber (Eds.), Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL 2013, Sofia, Bulgaria, August 8-9, 2013, Association for Computational Linguistics, 2013, pp. 2–11. URL: <https://aclanthology.org/W13-2402/>.
- [11] Y. Sari, M. Stevenson, A. Vlachos, Topic or style? exploring the most useful features for authorship attribution, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, Association for Computational Linguistics, 2018, pp. 343–353. URL: <https://aclanthology.org/C18-1029/>.
- [12] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, *Procesamiento del Lenguaje Natural* 69 (2022) 289–299.
- [13] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism (2021).
- [14] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.
- [15] I. D. Dinov, Data science and predictive analytics, Cham, Switzerland (2018).