

Unlocking Sentiments: Exploring the Power of NLP Transformers in Review Analysis

Javier Alonso-Mencia^{1,*}

¹University Carlos III of Madrid (UC3M), Madrid, Spain

Abstract

Sentiment analysis, a highly coveted area within Natural Language Processing, holds significant business potential by enabling the analysis of opinions across various textual forms. The article presents the participation of the Javier Alonso-Mencia team in the REST-MEX@IberLef 2023 Sentiment Analysis track. The primary objective was to predict the polarity of tourists' opinions as well as identifying the country of origin and the type of tourist attraction in reviews written in Spanish.

To address this task, the author employed fine-tuned Transformers, specifically designed for sentiment analysis. In addition, data balancing techniques were utilized to enhance the model's performance. Through a series of comprehensive experiments, the findings revealed that the fine-tuned transformer models delivered remarkable results, securing a notable second-place ranking in the REST-MEX@IberLef 2023 shared task.

Keywords

Sentiment analysis, REST-MEX, 2023, Transformers, NLP, Spanish,

1. Introduction

Sentiment analysis, a crucial aspect of NLP, involves discerning opinions and their polarity. It aids organizations in efficiently extracting opinions and identifying their source country and the type of place being reviewed [1]. The tourism sector greatly benefits from sentiment analysis in Spanish-speaking countries, particularly in countries like Colombia, Mexico, and Cuba, where tourists actively share their experiences on social media platforms [2].

Mexico heavily relies on tourism for economic growth and employment [3]. However, the COVID-19 pandemic has adversely affected global tourism, posing challenges for developing economies. To recover, it is essential to enhance productivity in sectors like tourism. Leveraging platforms such as Tripadvisor, with their wealth of text-based data, enables us to understand tourist preferences [4, 5].

IberLEF 2023, September 2023, Jaén, Spain


*Corresponding author.

✉ javilonso9@gmail.com (J. Alonso-Mencia)

🌐 <https://javieralonso.io/> (J. Alonso-Mencia)



© 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)

Recent advancements in NLP, particularly Transformers, have revolutionized sentiment analysis by comprehending context and delivering superior results [6]. This paper proposes using Transformers for sentiment analysis in the REST-MEX@IberLef 2023 shared task, focusing on tourist satisfaction and classifying the type of place (restaurant, hotel, or tourist attraction) [7]. By combining sentiment analysis with the identification of source countries, this competition aims to provide comprehensive insights into the sentiment towards various establishments in Colombia, Mexico, and Cuba.

Sentiment analysis plays a vital role in understanding opinions [8, 9]. By leveraging advancements in NLP techniques like Transformers, this competition seeks to achieve state-of-the-art sentiment analysis results and extract valuable insights from vast text data [10, 11, 12].

2. Methodology

2.1. Data analysis

The collection of labeled data given in the competition consists of a total of 251,702 comments obtained from tourists who shared their opinion on TripAdvisor between 2002 and 2022. Each comment contains the following fields: title, review, attraction (type of place), polarity (1-5) and country (Colombia, Cuba, Mexico).

First, it was required to identify if the training dataset was well-balanced, that is, all classes have similar number of instances. An unbalanced dataset means that the models will not be able to train properly on the least represented classes, which might hinder the inference as there is not enough insight to classify them.

The analysis of polarity distribution on Figure 1 shows that polarity is highly biased towards the most positive polarity (value 5), representing 60% of the polarity values, meanwhile the lowest polarity (value 1) has a representation of about 2%.

In terms of the type of place (attraction) and country, it is more evenly distributed with "Attractive" and "Mexico" holding almost 50% respectively of the values and the other half divided between the other classes.

Table 1 presents a detailed analysis of the number of tokens per instance in the Title and Review fields. The table reveals that there are a total of 251,603 non-empty instances for the Title field and 251,700 non-empty instances for the Review field. The mean number of tokens for the Title field is 24, while for the Review field, it is 362. The standard deviation for the Title field is 15, indicating a relatively narrow distribution of token counts. In comparison, the standard deviation for the Review field is higher at 429, suggesting a wider variation in token counts. The minimum number of tokens for the Title field is 1, whereas for the Review field, it is 7. Finally, the maximum number of tokens observed in the Title field is 185, while in the Review field, it reaches as high as 20,438.

A preprocessing step was performed to remove instances which include empty titles, empty reviews or reviews with more than 5,000 characters. In total 411 long instances were removed.

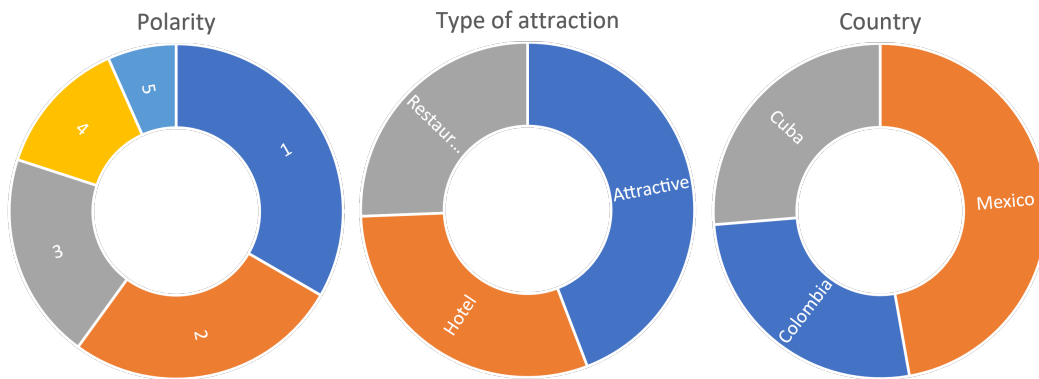


Figure 1: Instances distribution per class in different attributes.

Token analysis			
Title		Review	
Count	251603	Count	251700
Mean	24.2	Mean	362.8
Std.	15.71	Std.	429.7
Min.	1.0	Min.	7.0
25%	13.0	25%	175.0
50%	20.0	50%	250.0
75%	31.0	75%	401.0
Max.	185.0	Max.	20438.0

Table 1
Analysis of tokens in Title and Review fields

2.2. Data Balancing

To mitigate the significant class imbalance present in the original dataset, a data balancing strategy was implemented to improve the representation of minority classes. The initial distribution exhibited a substantial disparity, with Polarity class 5 dominating the majority of instances (157,095), followed by classes 4, 3, 2, and 1, which progressively decreased in count. The objective was to achieve a more equitable distribution while ensuring that no class exceeded 50,000 samples after the balancing process.

To accomplish this, a custom oversampling function was developed. The function took the dataset as input, along with the column indicating the sentiment label. Firstly, it computed the count of examples per label to quantify the existing class imbalance. Then, the maximum number of samples to add per class was determined.

The oversampling procedure involved iterating over each label individually. For each label, the function randomly selected examples from the corresponding class's dataframe, considering the number of instances required to reach the maximum sample count. The chosen examples were then appended to the original instances of the respective label. This process was repeated

for every class, ensuring that each label had a proportionate representation in the balanced dataset.

After oversampling, the resulting dataset was shuffled to introduce randomness and prevent any order-based bias. The final dataset exhibited a more balanced distribution compared to the original data. Table 2 shows the new data distribution after performing the custom oversampling.

After careful consideration, alternative techniques such as data augmentation through summarization were evaluated for addressing the class imbalance. However, based on previous findings [13], it was determined that this approach did not yield significant improvements in the performance of the sentiment analysis model. Therefore, the decision was made to exclude data augmentation through summarization from the current approach.

Polarity		Type of attraction		Country	
Class	Num. instances	Class	Num. instances	Class	Num. instances
5	156854	Attractive	131983	Mexico	148707
4	64848	Hotel	106029	Colombia	93904
3	60118	Restaurant	81917	Cuba	77318
2	20820				
1	17289				

Table 2

Number of instances for Polarity, Attraction and Country classes after data oversampling

2.3. Create dataset

In the dataset preparation phase, the columns of the dataset were processed to ensure compatibility with the sentiment analysis task. First, a code snippet was executed to convert the "Polarity" column values from a range of 1 to 5 to a more standardized range of 0 to 4. Similarly, the "Type" column values were transformed, assigning the value of 0 to "Hotel," 1 to "Restaurant," and 2 to "Attractive." Furthermore, the "Country" column values were modified, mapping "Mexico" to 0, "Cuba" to 1, and "Colombia" to 2.

To facilitate the analysis, a new column called "Title_Review" was created by concatenating the "Title" and "Review" columns. This combined column would capture both the concise titles and the comprehensive reviews, providing a comprehensive representation of the text data.

Subsequently, the dataset was shuffled to ensure the randomness of instance ordering. For the experimental setup, 10% of the instances were set aside as a test set, while the remaining 90% constituted the training set. The resulting dataset was organized into two subsets: the "train" subset, comprising 287,936 instances, and the "test" subset, comprising 31,993 instances.

The prepared dataset [14], consisting of columns such as "Title_Review", "Polarity", "Country", and "Type", was now ready for further analysis and model training.

2.4. Training

2.4.1. Transformers

In recent years, significant progress has been made in leveraging Transformers for various natural language processing (NLP) applications. This progress has been facilitated by the availability of platforms like Huggingface [15], which offer convenient means to utilize and train pretrained models. By employing pretrained models, transfer learning can be harnessed, allowing the model to benefit from pre-existing knowledge and thereby reducing training time and resource requirements while enhancing performance [16]. Huggingface models play a pivotal role in simplifying this process as they offer ready-made implementations for diverse NLP tasks.

Two different transformers approaches based on RoBERTa [17] were followed. RoBERTa is an optimized version of the BERT model, focusing on the encoder part of the transformer architecture [18, 19]. It employs masked language modeling during training, masking around 15% of the tokens. This makes RoBERTa well-suited for tasks that require sentence-level understanding and informed decision-making.

- **PlanTL-GOB-ES/roberta-base-bne** [20]: This model is a variant of RoBERTa specifically trained with a Spanish vocabulary. It utilizes the same tokenizer as the original RoBERTa model for consistent tokenization.
- **cardiffnlp/twitter-xlm-roberta-base** [21]: This model is a variant of RoBERTa trained on 198M multilingual tweets from Twitter. It is a multilingual model.

2.4.2. Models trained

The different models trained are presented in this section. The problem to solve was divided into 3 categories depending on which label was required to be predicted. Therefore, at least one model per attribute (Polarity, Attraction and Country) was trained.

Polarity

Model 1 - Polarity

- cardiffnlp/twitter-xlm-roberta-base
- 4 epochs
- lr: 2e-5
- batch size: 16

Model 2 - Polarity

- cardiffnlp/twitter-xlm-roberta-base
- 7 epochs
- lr: 2e-5
- batch size: 16

Model 3 - Polarity

- PlanTL-GOB-ES/roberta-base-bne
- 2 epochs
- lr: 2.5e-5
- batch size: 16

Model 4 - Polarity

- PlanTL-GOB-ES/roberta-base-bne
- 3 epochs
- lr: 2.5e-5
- batch size: 16

Model 5 - Polarity

- cardiffnlp/twitter-xlm-roberta-base
- 8 epochs
- lr: 1e-5
- batch size: 16

Type of attraction

Model 1 - Attraction

- cardiffnlp/twitter-xlm-roberta-base
- 4 epochs
- lr: 1e-5
- batch size: 16

Country

Model 1 - Country

- cardiffnlp/twitter-xlm-roberta-base
- 4 epochs
- lr: 1e-5
- batch size: 16

Model 2 - Country

- cardiffnlp/twitter-xlm-roberta-base
- 4 epochs
- lr: 2e-5
- batch size: 16

3. Results and discussion

3.1. Evaluation

The evaluation section assesses the performance and effectiveness of the developed models in addressing the research objectives. This section presents the evaluation metrics used to measure the models' performance.

The evaluation results in Table 3 indicate the performance of different models trained for each class. For the "Polarity" classification task, Model 1.2 achieved the highest macro F1 score of 0.8461, outperforming other models. In the "Type of attraction" classification, Model 2.1 demonstrated the highest macro F1 score of 0.9941, showcasing superior performance. For the "Country" classification, Model 3.2 achieved the highest macro F1 score of 0.9566, indicating its effectiveness. These findings underscore the significance of selecting appropriate models tailored to specific classification tasks, ultimately enhancing the overall performance of the NLP system.

Polarity			Type of attraction			Country		
Model	Macro F1 score	Val. loss	Model	Macro F1 score	Val. loss	Model	Macro F1 score	Val. loss
1.1	0.8217	0.4959	2.1	0.9941	0.1232	3.1	0.9375	0.2714
1.2	0.8461	0.5889				3.2	0.9566	0.2185
1.3	0.8186	0.4623						
1.4	0.8508	0.6224						
1.5	0.6070	0.7781						

Table 3

Evaluation results for each trained model on validation dataset

Due to the fact that it was possible to submit any number of models, it was decided to send all combinations of models in order keep the models with better results. Table 4 indicates the 10 submissions created for the competition.

3.2. Competition results

The sentiment analysis competition showcased impressive results, with the 6th submission, out of a total of 10 submissions, emerging as the second-best performing entry in the competition (see Table 5). The team's submission comprised three models tailored for polarity, attraction, and country classes respectively.

Submission	Polarity model	Attraction model	Country model
1	1.1	2.1	3.1
2	1.1	2.1	3.2
3	1.2	2.1	3.1
4	1.2	2.1	3.2
5	1.3	2.1	3.1
6	1.3	2.1	3.2
7	1.4	2.1	3.1
8	1.4	2.1	3.2
9	1.5	2.1	3.1
10	1.5	2.1	3.2

Table 4
Models included in each of the submissions for the competition

The evaluation metrics demonstrated the effectiveness of the team’s approach in sentiment analysis. The Sentiment Track Score achieved a noteworthy value of 0.766, highlighting the team’s competence in accurately predicting sentiment. The macro F1 scores for each class were remarkably high, with polarity achieving 0.602, attraction achieving 0.988, and country achieving 0.936.

The extraction of the polarity class proved to be more challenging compared to the country or type of attraction classes. Several factors contributed to this difficulty. Firstly, the inherent subjectivity of sentiment analysis poses challenges in accurately capturing the nuanced polarity of textual data. Additionally, the presence of ambiguous or sarcastic language further complicates the task of polarity classification. Furthermore, the variability and diversity of expressions used to convey sentiment within the polarity class create additional complexity. These factors collectively contributed to the increased difficulty in extracting polarity compared to the country or type of attraction, highlighting the intricacies involved in accurately discerning sentiment from textual data.

Rank	Run	Sentiment Track Score	Macro F1 (Polarity)	Macro F1 (Type)	Macro F1 (Country)
1st	LKE-IIMAS-Team_RUN_2	0,779	0,621	0,990	0,942
2nd	javier_alonso-Team_sentiment_sub6	0,766	0,602	0,988	0,935
3th	IIMAS-UNAM-Team_resultados	0,750	0,593	0,979	0,902

Table 5
Final results from Sentiment Analysis competition Rest-Mex 2023

4. Conclusions

In this paper, the application of transformer-based models for sentiment analysis in the context of an NLP competition was explored. Two models based on RoBERTa, *PlanTL-GOB-ES/roberta-base-bne* and *cardiffnlp/twitter-xlm-roberta-base*, were employed to solve the task. The results obtained from the evaluation indicate that strong performance was demonstrated this approach across multiple evaluation metrics.

It was observed that the 6th submission, consisting of three models for polarity, type of attraction, and country classes, achieved the second-best result in the competition.

Furthermore, it was noted that the extraction of the polarity class presented greater challenges compared to the country and type of attraction classes. The increased complexity in accurately discerning polarity from textual data was attributed to the inherent subjectivity of sentiment analysis, combined with the presence of ambiguous language and diverse expressions.

As future work to enhance the performance of sentiment analysis models, further exploration of preprocessing techniques and data balancing methods can be considered. The application of more advanced preprocessing techniques, such as lemmatization, stemming, or part-of-speech tagging, could help improve the quality of the input data by reducing noise and capturing more meaningful features.

The experimentation can be found in a Github repository [22].

References

- [1] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
- [2] Y. Qu, Y. Mao, J. Chen, J. Li, *Microblogging after a major disaster in china: A case study of the 2010 yushu earthquake*, in: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 2011.
- [3] S. de Turismo (SECTUR), *Anuario estadístico de turismo en méxico 2019*, 2019.
- [4] V. Kumar, A. R. Singh, M. K. Tiwari, *Tourism forecasting models: A review of literature*, *Tourism Management Perspectives* 19 (2016) 39–55.
- [5] 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>.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention is all you need*, *Advances in Neural Information Processing Systems* 30 (2017) 5998–6008.
- [7] 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).
- [8] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-

- González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, *Procesamiento del Lenguaje Natural* 67 (2021).
- [9] 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).
- [10] B. Pang, L. Lee, Opinion mining and sentiment analysis, in: *Foundations and Trends in Information Retrieval*, volume 2, 2008, pp. 1–135.
- [11] 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>.
- [12] 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. URL: <https://doi.org/10.1080/0952813X.2022.2153276>. doi:10.1080/0952813X.2022.2153276. arXiv:<https://doi.org/10.1080/0952813X.2022.2153276>.
- [13] M. P. Enríquez, J. A. Mencía, I. Segura-Bedmar, Transformers approach for sentiment analysis: Classification of mexican tourists reviews from tripadvisor (2022).
- [14] Javier Alonso, *rest23_sentiment_data_v3_oversampling* (revision 4c3329a), 2023. URL: https://huggingface.co/datasets/javilonso/rest23_sentiment_data_v3_oversampling. doi:10.57967/hf/0675.
- [15] Hugging face – the ai community building the future., 2023. URL: <https://huggingface.co/>.
- [16] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, *Proceedings of the IEEE* 109 (2020) 43–76.
- [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [18] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR abs/1810.04805* (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [20] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022). URL: <https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley>. doi:10.26342/2022-68-3.
- [21] F. Barbieri, L. E. Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, 2022. arXiv:2104.12250.

[22] J. Alonso, Restmex23_nlp, https://github.com/javilonso/RestMex23_NLP, 2023. Accessed: May 23, 2023.