CEUR-WS.org/Vol-3496/restmex-paper18.pdf

Analyzing Sentiment, Attraction Type, and Country in Spanish Language TripAdvisor Reviews Using Language Models

Pedro Mirabal[1],*, Suilen Hernández-Alvarado[2] and José Ignacio Abreu Salas[3]

[1]*Departamento de Ingeniería Informática. Unversidad Católica de Temuco. Chile.*
[2]*Database Laboratory, Universidade da Coruña. Spain.*
[3]*U.I. for Computer Research. University of Alicante. Spain.*

Abstract

This paper describes our participation in the Rest-Mex 2023 Sentiment Analysis Task. We proposed an ensemble of (i) a cascade of transformer-based two-class classifiers biased to lowering the Mean Average Error in Polarity, and (ii) multi-class transformer-based classifiers for the prediction of the Type and Location of the messages. Our system achieved a sentiment track score of 0.719.

Keywords

Sentiment Analysis, Deep Learning, Transformer Models

1. Introduction

Sentiment Analysis, a subfield of Natural Language Processing (NLP), enables the examination of an individual's opinions towards various entities, including services and products, by categorizing them into distinct classes. These classes can encompass positive, negative, neutral, or even more nuanced gradations. This particular task has garnered significant interest due to the potential for stakeholders to utilize data obtained from social media platforms and specialized websites like Tripadvisor, empowering them to make informed decisions based on data-driven insights. Nonetheless, several challenges persist, such as the disparate availability of linguistic resources across different languages [1]

To advance the field of Sentiment Analysis, several challenges have been established to foster research and development in this area. Notably, initiatives like SemEval, which commenced with its first edition in 2007, have played a crucial role in this regard. Additionally, other challenges such as IberLEF have contributed significantly to the advancement of Sentiment Analysis. More recently, a noteworthy challenge known as Rest-Mex has emerged, further enriching the landscape of sentiment analysis research and providing opportunities for researchers and practitioners to explore and address the complexities associated with sentiment classification

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.
pedro.sanchez@uct.cl (P. Mirabal); s.hernandez@udc.es (S. Hernández-Alvarado); ji.abreu@ua.es (J. I. A. Salas)
0000-0001-7345-6007 (P. Mirabal); 0000-0002-4887-9216 (S. Hernández-Alvarado); 0000-0002-4637-4206 (J. I. A. Salas)

© 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)

tasks [10, 11, 12, 13].

In recent times, the field of Sentiment Analysis has witnessed notable advancements through the utilization of Deep Learning techniques. A comprehensive exploration of this subject can be found in a survey entitled "Deep learning for sentiment analysis: A survey" by Zhang et al. (2018) [2]. In a similar vein, several research teams have leveraged similar strategies and participated in competitions pertaining to Sentiment Analysis, yielding commendable outcomes[3, 4, 5].

It is worth noting the results obtained by the winners of the RestMex 2022 edition, where the winning team [6] utilized a large collection of attributes computed with the UMUTextStats tool combined with models based on BERT [16] and RoBERTa [17]. They achieved a score of 0.892, which was very close to the score obtained by the second team. In the case of the second place [8], the most relevant part of their work is the meticulous preprocessing of the data they carried out. They removed duplicate instances from the training dataset and translated any opinion that was not in Spanish. If the translation was not possible, they discarded that instance. The second-place team achieved their best result by using a variant of BERT, similar to the first-place team.

In this paper, we further explored the proposal of a cascade of biased two-class classifiers for predicting Polarity described in [9]. We ran a comparative study of different language models in Spanish. We aim to provide a detailed account of our involvement in the Rest-Mex 2023 Sentiment Analysis Subtask [13] which primarily focuses on sentiment classification, wherein the objective is to develop systems capable of accurately predicting the Polarity, the Location, and the Type of attraction of tourist opinions concerning various locations in Mexico, Cuba, and Colombia. The report is structured as follows. Section 2 describe the training dataset. In section 3 we present details of our approach. Section 4 is devoted to the analysis of the results.

2. Task and Data Description.

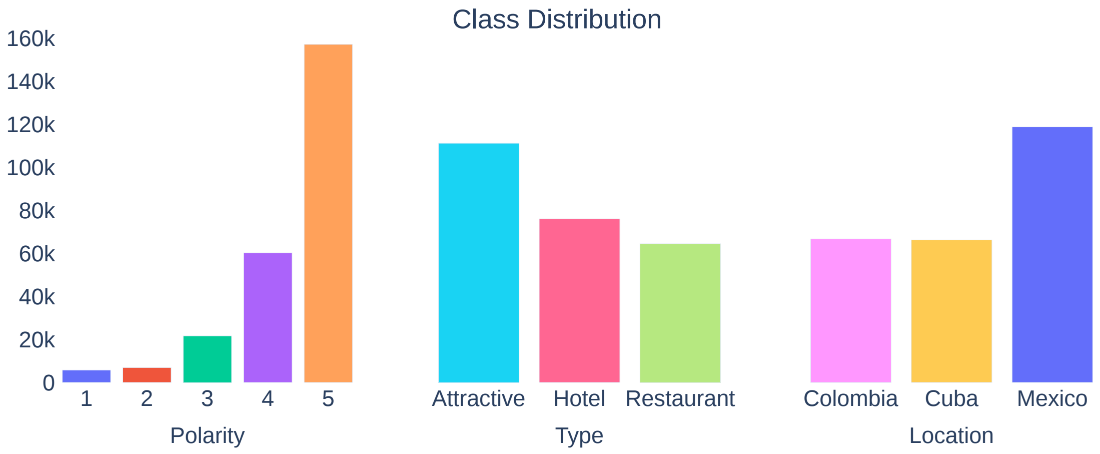


Figure 1: Class distribution – number of instances – of the training data.

In this section, we describe the data provided by the organizers for this subtask and its characterization. The corpus consists of 251.702 opinions. Each opinion is classified as an integer, between [1, 5], where 1 represents the most negative polarity and 5 the most positive. For each opinion, organizers also provided information about Location [Colombia, Cuba, Mexico] and Type [Attractive, Hotel, Restaurant]. The organizers split the corpus $70\% - 30\%$ approximately. 70% of the data was delivered to the participants with complete information about each opinion, and 30% was reserved for the final testing of competing models.

Figure 1 shows the class distribution for Polarity, Type and Location. Analyzing the representation of each of the classes, we detected that they had a high level of imbalance, with class 5 as the majority class, with a total of $157,095$ instances, representing 62.41% of the total, a great contrast with the class 1, for which only $5,772$ instances were provided, for the 2.21%. The presence of the rest of the classes is as follows: $60,227$ instances for class $4, 21,656$ instances for class 3 and $6,952$ instances for class 2, each representing 30.71%, 13.21% and 2.98% respectively. On the other hand, regarding the Location, the distribution is as follows: $66,703$ represents Colombia, $66,223$ represents Cuba, and $118,776$ for Mexico. While the data imbalance in terms of location is not as significant as in the case of polarity, there is still an over-representation of instances from Mexico. Finally, in terms of attraction type, there are $76,042$ instances categorized as Hotel, $64,472$ instances categorized as Restaurant, and the remaining $111,188$ instances categorized as Attractive.

3. System Description.

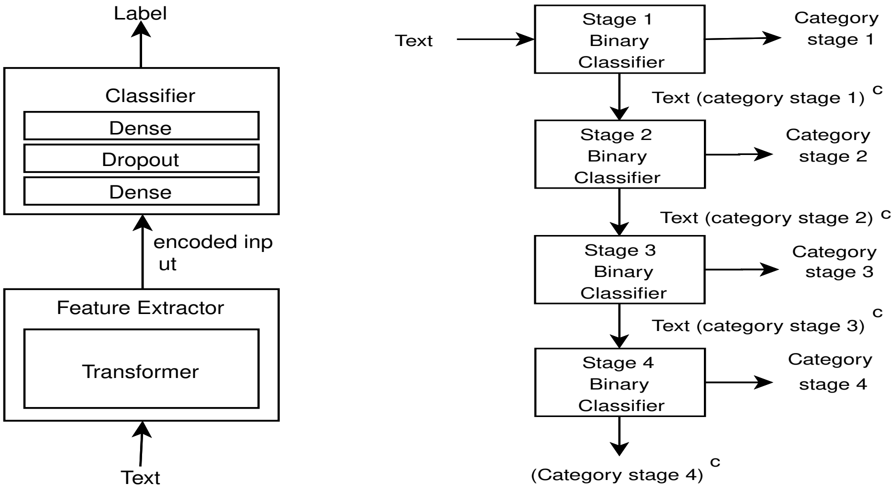
Our approach (https://github.com/joseias/2023-rest-mex) implements an ensemble of three classifiers for the multi-label classification problem posed by the challenge. We have separated classifiers for the Polarity, the Type, and the Location of the review.

For Type and Location, we use the architecture depicted in 2, as implemented by the [Bert|Roberta]ForSequenceClassification modules in Huggin Face (https://huggingface.co). This architecture has been used for the task by other authors such as [7] and [6].

In the case of the Polarity, our approach further studies the proposal of the cascade of biased classifiers described in [9]. The architecture is described in 2b. The main hypothesis is building classifiers to learn the target class but with a bias towards the other classes with lower miss-classification costs. That is, in case of a miss-classification, the model is biased towards other of the better options. For example, if the target is of polarity of 1, the error with lower cost in terms of Mean Square Error (MSE) is to assign a polarity of 2. In the next section, we delve into the details of this approach.

3.1. Cascade of Biased Two-class Classifiers.

As was commented before, for Polarity we leveraged the second proposal of [9], an ensemble of biased two-class classifiers arranged as shown in Fig. 2b. The ensemble is comprised of four classifiers. The classifier at each stage is trained to separate the instances from the original 5 possible polarities into two categories, (i) the target for the stage i, let us call it C_i, (ii) the others classes C_i^c. Both, C_i and C_i^c are defined as biasing the classifier towards classes with low miss-classification costs with respect to MSE.


Label
Classifier
Dense
Dropout
Dense
encoded inp ut
Feature Extractor
Transformer
Text
Text
Stage 1 Binary Classifier
Category stage 1
Text (category stage 1) c
Stage 2 Binary Classifier
Category stage 2
Text (category stage 2) c
Stage 3 Binary Classifier
Category stage 3
Text (category stage 3) c
Stage 4 Binary Classifier
Category stage 4
(Category stage 4) c

(a) BERT-based multi-class.

(b) Cascade classifier.

Figure 2: Architectures of (a) the multi-class classifier for Type and Location, (b) the cascade ensemble.

For example, for the binary classifier at stage 1, $C_1 = \{1\}$ and $C_1^c = \{2, 3, 4, 5\}$. This classifier is learning to tear apart instances with a Polarity of 1 and instances with other polarities. For stage 2, the sets are defined as $C_2 = \{1, 2\}$ and $C_2^c = \{3, 4, 5\}$. This classifier is biased to classify as 2 the instances with a Polarity of 1. At the inference step, this classifier is used after the classifier at stage 1. Thus, if the later model miss-classifies a true 1 example, the hypothesis is that the model at stage 2 will classify this instance as 2, contributing to minimizing MSE. The values of C_i and C_i^c are set in a similar fashion for stage 3. And, in the step 4 they are $C_4 = \{5\}$ and $C_4 = \{1, 2, 3, 4\}^c$. It is worth noting that no model for Polarity of 4 is trained. This class is inferred by the other models.

To infer the class of an instance, among the original 5 categories, the example is sent to the model at stage 1. If classified as C_1 we assign Polarity of 1. In case of being classified as C_1^c, the example is sent to the model at step 2 following the same procedure. If finally the instance is classified as C_4^c, i.e. it reached the last stage, we assign Polarity of 4.

4. Experiments and Results.

As a contribution over [9], in this work we aim to compare the performance of different language models for Spanish when fine-tuned for the task. It is worth noticing that the models we studied are not the only available. We prioritized them considering they are well documented, they used different corpus for pre-training, are BERT or RoBERTa based models, and also the time we had to run experiments. In all cases, we choose the base cased version of the model. We trained cascade-based models for Polarity, as well multi-class models for Type, Location and also for Polarity.

In our experiments, we use as input the concatenation of the title and the content of the review, together with the separator token. Training set was split stratified in sets for $training$

(90%, development (*dev*) (10%) and *test* (10%). We used the *train* and *dev* sets to fine-tune each model. The best checkpoint was the one with higher balance accuracy, and in cases of very small differences (less than 0.001), the lower loss in *dev* was the selection criteria.

Table 1 summarizes some details about the models. For the fine-tuning, we set hyperparameters such as learning rate, or weight decay as reported by the authors, except for the batch size (32) and the number of training epochs (5). When to our knowledge the authors didn't report the values, defaults TensorFlow 2.11 were used. As a training algorithm, we use AdamW with a linear learning rate decay scheduler. In evaluation time, we chose the best checkpoint considering the balance accuracy metric since for Polarity the data set is highly unbalanced.

Table 1
Spanish Language Models evaluated. In Hyperparameters, values that are different from the defaults for AdamW optimizer. Columns Polarity (Pol.) Stage 1 to Stage 4 indicates the epoch of the best checkpoint for balance accuracy over the dev set for the cascade-based models. Similar for the Polarity (Pol.), Type, and Location (Loc.) columns.

No	Model	Hyper parameters	Best checkpoint from 5 epoch						
			Pol. Stage 1	Pol. Stage 2	Pol. Stage 3	Pol. Stage 4	Pol.	Type	Loc.
1	BERTIN [21]	learning rate = 5e-5	2	5	5	5	4	5	4
2	BETO [20]	weight decay=0.01 learning rate = 2e-5	3	1	1	2	1	4	3
3	MarIA-BNE [19]	weight decay=0.1 learning rate = 1e-5	1	1	2	1	3	4	3
4	RoBERTuito [22]	learning rate = 5e-5	2	2	1	5	5	5	2

Examining Table 1, we observed that most models degraded their performance after 1 or 2 epoch, except for BERTIN and RoBERTuito. This is a signal that they might need more hyperparameter tuning, in particular BERTIN. The cascade models at stages 1, 2, and 4 achieved 0.00 recall, i.e. all instances were classified as class 3 or 5.

Table 2 shows the Macro F1 for Polarity, Type, and Country for each classifier over the test set we separated. From our experiments, we observed some interesting facts. It seems BERTIN and RoBERTuito performance degrades notably in the presence of class imbalance, at least for our experimental setup. This is not surprising in the case of RoBERTuito since the model was trained only with Twitter data. However, the poor results from BERTIN in the cascade classifier were not expected, deserving further studies. On the other hand, BETO and MarIA achieved very discrete results, but in line with the challenge data, where our BETO model achieved 0.515

For Type, all systems except RoBERTuito achieved results of about 0.99. Given the global results of the task, it seems this problem is not as hard as the Polarity. The same conclusion holds for the Country classification. It is worth noting that in this case, RoBERTuito managed to achieve an F1 Macro of 0.93.

Table 2
Macro F1 for the different models and sub tasks.

Model	Macro F1 Polarity (Cascade)	Macro F1 Polarity	Macro F1 Type	Macro F1 Country
BERTIN	0.213	0.077	0.985	0.930
BETO	0.451	0.594	0.988	0.942
MarIA-BNE	0.417	0.587	0.988	0.936
RoBERTuito	0.319	0.486	0.979	0.934

5. Conclusions and Future Work

In this paper, we have described the model proposed by UCT-UA in the Sentiment Analysis subtask at Rest-Mex 2023. Also, we studied the performance of different language models in Spanish. The study is in a very preliminary stage, thus is better to interpret observed results with caution.

The results in our primary submission were obtained from the model described in the Results section as BETO, this model ranked $5th$ out of 17 submissions, achieving 0.719 for the Sentiment Track Score. Globally, it seems the cascade strategy designed to lower MAE does not perform well with the Sentiment Track Score, at least with our experimental setup and the degree of imbalance of the data. This is consistent with the results of [7] and [9] where fine-tuned BERT model achieved better results than the cascade-based approach.

As future work, we are interested in further studying the performance of the different models, to draw more sound conclusions. In particular, those trained with general Spanish corpus, or others from specialized domains such as review data. Also, it would be interesting to evaluate multilingual models.

6. Acknowledgments

This research has been funded by: the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport), through the project NL4DISMIS: Natural Language Technologies for Dealing with dis- and misinformation (CIPROM/2021/021); MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/ PRTR through the project ClearText $< TED2021-130707B-I00 >$.

References

[1] Agüero-Torales, M., Abreu-Salas, J., López-Herrera, A.: Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, vol 107 (2021)

[2] Zhang, L., Wang, S. and Liu, B.: Deep learning for sentiment analysis: A survey. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. vol 8, number 4 (2018)

[3] González, J.A., Hurtado, L.F. and Pla, F. : ELiRF-UPV at TASS 2019: Transformer Encoders for Twitter Sentiment Analysis in Spanish. In: Proc. of IberLEF@ SEPLN, (2019)

[4] Pastorini, M., Pereira, M., Zeballos, N., Chiruzzo, L., Rosá, A. and Etcheverry, M.: RETUYT-InCo at TASS 2019: Sentiment Analysis in Spanish Tweets. In: Proc. of IberLEF@ SEPLN, (2019)

[5] González, J., Pla, F. and Hurtado, L.: ELiRF-UPV at SemEval-2017 Task 4: sentiment analysis using deep learning. In: Proceedings of the 11th international workshop on semantic evaluation SemEval-2017, (2017)

[6] García-Díaz, J., Rodríguez-García, M., García-Sánchez, F. and Valencia-García, R.: UMUTeam at REST-MEX 2022: Polarity Prediction using Knowledge Integration of Linguistic Features and Sentence Embeddings based on Transformers. (2022)

[7] Vásquez, J. and Gómez-Adorno, H. and Bel-Enguix, G..: Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor, pages 165–170, (2021).

[8] Ramírez, S., Daniel, D. and Bedmar, I.: Recommendation System Rest-Mex 2022 for Mexican Tourism Using Natural Language Processing. 2022

[9] Abreu, J., Mirabal, P., Ballester-Espinosa, A.: Cascade of Biased Two-class Classifiers for Multi-class Sentiment Analysis. on Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September 2021, CEUR Workshop Proceedings, vol 2943, pages 185–191, 2021.

[10] Álvarez-Carmona, Miguel Á and Aranda, Ramón and Arce-Cárdenas, Samuel and Fajardo-Delgado, Daniel and Guerrero-Rodríguez, Rafael and López-Monroy, A. Pastor and Martínez-Miranda, Juan and Pérez-Espinosa, Humberto and Rodríguez-González, Ansel: Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism. Procesamiento del Lenguaje Natural, vol 67 (2021)

[11] Álvarez-Carmona, Miguel Á and Díaz-Pacheco, Ángel and Aranda, Ramón and Rodríguez-González, Ansel Y and Fajardo-Delgado, Daniel and Guerrero-Rodríguez, Rafael and Bustio-Martínez, Lázaro; Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts. Procesamiento del Lenguaje Natural, vol 69 (2022)

[12] Álvarez-Carmona, Miguel Á and Aranda, Ramón and Guerrero-Rodríguez, Rafael and Rodríguez-González, Ansel Y and López-Monroy, A Pastor; A Combination of Sentiment Analysis Systems for the Study of Online Travel Reviews: Many Heads are Better than One. Computación y Sistemas, vol 26, 2022.

[13] Álvarez-Carmona, Miguel Á and Díaz-Pacheco, Ángel and Aranda, Ramón and Rodríguez-González, Ansel Y and Bustio-Martínez, Lázaro and Muñiz-Sánchez, Victor and Pastor-López, A Pastor and Sánchez-Vega, Fernando: Overview of Rest-Mex at IberLEF 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts. Procesamiento del Lenguaje Natural, vol 71 (2023)

[14] Calvo, H., Gambino, O.: Cascading classifiers for Twitter sentiment analysis with emotion lexicons. In: Proc. Int. Conf. on Intelligent Text Processing and Computational Linguistics, pp. 270-280. (2016)

[15] Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: Proc. of PML4DC at ICLR. (2020)

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional

transformers for language understanding. (2018)

[17] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019)

[18] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1631-1642. (2013)

[19] Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pámies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., & Villegas, M. (2022). MarIA: Spanish Language Models. Procesamiento del Lenguaje Natural, 68, pages 39-60, (2022)

[20] Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. Spanish pre-trained bert model and evaluation data. Pml4dc at iclr. (2020)

[21] De la Rosa, J., Ponferrada, E. G., Romero, M., Villegas, P., de Prado Salas, P. G., & Grandury, M. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. Procesamiento del Lenguaje Natural, 68, 13-23. (2022)

[22] Pérez, J. M., Furman, D. A., Alemany, L. A., & Luque, F. M. (2022, June). RoBERTuito: a pre-trained language model for social media text in Spanish. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7235-7243).