DCI-UG participation at REST-MEX 2021: A Transfer Learning Approach for Sentiment Analysis in Spanish

Geovanni Velazquez Medina and Delia Irazú Hernández Farías

División de Ciencias e Ingenierías Campus León,
Universidad de Guanajuato
[g.velazquezmedina|di.hernandez]@ugto.mx

Abstract. In this paper, we describe the system proposed by the DCI-UG team for participating in the REST-MEX shared task, concretely in the Sentiment Analysis subtask. The proposed method is based on a modified Spanish BERT-base architecture model. Two different settings were evaluated for dealing with Spanish Sentiment Analysis. The obtained results show that our method is able to identify the polarity degree of user opinions about touristic places, particularly the ones labeled with the highest strength. Our proposals ranked at the 4^{th} and 6^{th} according to the official results with a MAE value of 0.562 and 0.606, respectively.

Keywords: Sentiment Analysis · Transfer Learning · BERT-BETO

1 Introduction

Nowadays, users in social media tend to share their ideas, experiences, and opinions about practically any topic. Taking advantage of such data is very important for many domains: for example, it could serve for monitoring services, products, etc. Sentiment Analysis (SA) aims to identify the subjective information in a given piece of text. It has been one of the most active areas in Natural Language Processing (NLP) due to its importance beyond computer sciences such as business and industrial applications [1]. Sentiment Analysis can be applied at different levels: whole document, sentences, particular aspects of a given target, etc. Research in SA has been carried out from different perspectives ranging from rule-based approaches to traditional machine learning, and recently deep learning methods. Furthermore, it is strongly related to more complex tasks like emotions detection [2]. Despite the efforts carried out in this area, many challenges remain [3] such as, for example dealing with figurative language devices (like irony and sarcasm) [4].

IberLEF 2021, September 2021, Málaga, Spain.

However, notwithstanding being one of the more spoken languages around the world, literature concerning SA in Spanish is scarce. Most of the available tools and resources for performing such a complex task have been developed for the English language. Aiming to identify combinations of different text preprocessing techniques for improving the performance of traditional word-base combinations, the authors in [5] experimented with two different datasets of texts written in Spanish. Deep learning-based approaches have been also exploited for this task [6, 7]. Transfer learning has been also evaluated on SA in Spanish with competitive results [8]. As mentioned before, there is a lack of resources for performing SA in Spanish. In this sense, some works have attempted to develop resources in Spanish [9] while others have proposed to adapt methods and resources from English to Spanish [10]. Among the efforts for promoting the research on SA in Spanish, there is the TASS[1]: Workshop on Semantic Analysis at SEPLN, which includes a task dedicated to Sentiment Analysis on Spanish text as one of its evaluation campaigns. A more comprehensive review on Spanish Sentiment Analysis can be found in [11, 12].

In this paper, we describe our participation for the Spanish Sentiment Analysis subtask in the framework of the *Recommendation System for Text Mexican Tourism (REST-MEX)* shared task. The aim was to assign a polarity degree for user comments about touristic places in Guanajuato, Mexico. The proposed approach is based on a modification of the well-known BERT architecture. Different parameter settings were evaluated during development phase. Our systems show a competitive performance at the shared task. In the following sections, we describe our proposal as well as the obtained results.

2 DCI-UG participation at REST-MEX

2.1 Task Description

This year, in the framework of the IberLEF 2021, a task denoted as *Recommendation System for Text Mexican Tourism* has been organized. The task aims to promote the development of intelligent systems in tourism by considering TripAdvisor comments written in Spanish. The task is composed by two sub tasks: i)*Recommendation System*, which goal is to determine the satisfaction degree of a tourist will have when visiting a given place; and, ii)*Sentiment Analysis* which aims to determine the polarity (within range 1 up to 5) of an opinion about a Mexican touristic place. Organizers provided data for training and test purposes for each subtask. Participants were allowed to submit two different systems proposals without any restriction about data for training purposes. The Mean Absolute Error (MAE) was used as evaluation metric for ranking the submissions. More details about the task can be found on the task overview [13].

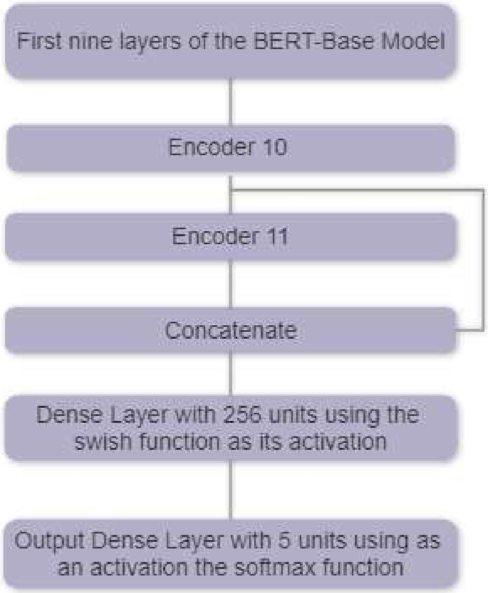
[1] http://tass.sepln.org/

2.2 Our Proposal

In order to tackle the Sentiment Analysis task, we propose a system based on the use of network-based deep transfer learning [14] with a Spanish BERT-Base uncased model [15]. We decided to exploit transfer learning for this task by considering the outstanding results showed by this approach in related tasks and also in order to deal with the limitations regarding amount of data available for training purposes. The model has an architecture which consists of 12 layers (also called transformer blocks) with 768 units and 12 self attention heads per layer [16]. The BERT-Base model was used since it has shown a better performance in accuracy compared to other models using the GLUE Test results as reference[17]. The BERT-Large has a better performance than the Base model but no instances of this model pre-trained with Spanish texts were found at the time the model was implemented [17].

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 with the purpose of being used as the input to a dense layer which has a swish activation function [18]. As final layer, a dense layer was used with five outputs (one for each class) using softmax as activation function. The complete modified architecture consisted of 13 layers of which 11 layers were kept from the original transformer blocks from the BERT-Base model and the last two were the dense layers added. Figure 1 shows a schematic representation of the modified BERT architecture.

Fig. 1. Modified BERT architecture used in the proposed approach


First nine layers of the BERT-Base Model
Encoder 10
Encoder 11
Concatenate
Dense Layer with 256 units using the swish function as its activation
Output Dense Layer with 5 units using as an activation the softmax function

As *baseline*, we exploited a Multi-Layer Perceptron model with four layers. As input to be fed, we used a Spanish embeddings model called *FastText embeddings from SUC*[2]. For all layers except to the last one, the Rectified Linear activation function worked as the output of each layer. The last layer has a standard softmax output.

The distribution of the official training data is as follows: 80, 145, 686, 1596, and 2690 for classes *pol-1*, *pol-2*, *pol-3*, *pol-4*, and *pol-5*, respectively. Attempting to compensate the imbalance distribution in the official data, we decided to increase instances of the minority classes. For doing so, we took advantage of the available training data for the sub task (i). We filtered out the instances labeled with a polarity rate equal to 1 and 2, these samples were added to the training data for Sentiment Analysis. From these subset, we identified some samples written in English, then we automatically translated them to Spanish using the *langdetect*[3] Python library.

In order to assess the performance of our proposal, once the test set was released, we decided to manually annotate a randomly selected subset of 200 instances. Three independent annotators carefully read each text and labeled it with a polarity degree. The inter-annotator agreement among them was 0.35 in terms of Fless' Kappa, which can be interpreted as a *fair agreement*. For experimental purposes, we used only those samples where the three annotators fully agreed. This left an annotated test set of 93 samples (hereafter this subset is denoted as *InHouseTest*). It is important to highlight that, the proportion of classes among this subset was similar to the official training set.

At the end, our models were trained with a total of 4562 samples, distributed as: 448, 634, 480, 1117, and 1883 for classes *pol-1*, *pol-2*, *pol-3*, *pol-4*, and *pol-5*, respectively, when the *InHouseTest* was not used as the validation set. In this case, the validation set consisted of 1560 samples taken only from the original training set distributed as: 24, 44, 206, 479 and 807 for classes *pol-1*, *pol-2*, *pol-3*, *pol-4*, and *pol-5*, respectively.

The modified BERT model was evaluated considering two optimizers: Adam [19] and LAMB [20]. With the former one, the best results were obtained with a batch size of 1, 2, 4, and 8. The average MAE obtained was around 0.47. On the other hand, when LAMB was used, the best MAE rate was around 0.457 with a batch size of 256. A bigger batch size could give better results[20], nonetheless in the experiments performed a greater batch size was not used due to hardware limitations. Table 1 shows the obtained results with different optimizers and learning rates when using the both settings of training data; in each case, the MAE and accuracy are presented. As it can be noticed, the same parameters configurations achieves the highest results in each dataset. Then, we decided to train our participating system with LABM optimizer and a learning rate of 0.00006.

[2] https://github.com/dccuchile/spanish-word-embeddings

[3] https://pypi.org/project/langdetect/

For our participation in the shared task, we decided to submit two different settings denoted as:

- *DCI-UG*$_a$. In this case, the model was trained with 70 percent of the training data (with data augmentation for classes 1 and 2) and using the remaining 30 percent of the data as a validation set.
- *DCI-UG*$_b$. This model was trained by using the whole training data (again including the additional data), and using the *InHouseTest* as the validation set to prevent model over-fitting.

Table 1. Best MAE obtained per optimizer considering both dataset settings.

Optimizer	MAE	Learning Rate	Accuracy
Official Training + Data augmentation			
Adam	0.464	0.000004	0.595
LAMB	0.457	0.00006	0.599
InHouseTest			
Adam	0.263	0.000003	0.812
LAMB	0.172	0.00006	0.870

Official Results

Our submissions ranked at the 4^{th} and 6^{th} according to the official results in the shared task for subtask (ii). In Table 2, we included the obtained results of our proposals as well as the ones obtained by the best ranked teams[4]: Minería UNAM and UCT-UA, and the Official Baseline (it was obtained by considering a Majority class approach). The official metric for evaluating the participating systems was the **MAE**. However, we decided also to include the results in terms of F-measure per class and in overall in each case.

In terms of the official evaluation metric, i.e., the MAE, our participating systems outperform the baseline and are better than the average score in the competition. The accuracy rate obtained in both cases is very competitive in comparison with the best ranked systems, even higher than the second place systems in the ranking. For what concerns to the overall F-measure, the rates obtained by our systems were considerably affected by the poor performance in the *pol-1* and *pol-2* classes. This could be due to the imbalance degree in the training data towards these two classes. It is also important to mention that, the obtained results of each of our systems in the official evaluation match with the ones obtained during development. The *DCI-UG*$_a$ ranked better than

[4] In the table we use the acronyms UNAM and UCT with a number in subscript indicating the corresponding system according to the official results. Further details on these approaches can be found in the task overview [13].

Table 2. Official results of the *Sentiment Analysis* subtask at REST-MEX 2021

Rank	Team	MAE	Accuracy	F-measure					
				Overall	*pol-1*	*pol-2*	*pol-3*	*pol-4*	*pol-5*
1	UNAM$_1$	0.475	56.723	0.428	0.222	0.307	0.474	0.443	0.692
2	UCT-UA$_2$	0.545	53.249	0.451	0.377	0.396	0.421	0.389	0.670
3	UCT-UA$_1$	0.561	53.835	0.403	0.298	0.254	0.374	0.403	0.686
4	**DCI-UG$_a$**	**0.562**	**53.339**	**0.287**	**0**	**0.083**	**0.312**	**0.344**	**0.694**
5	UNAM$_2$	0.582	54.783	0.242	0	0	0.247	0.25	0.717
6	**DCI-UG$_b$**	**0.606**	**53.700**	**0.253**	**0.125**	**0**	**0.328**	**0.102**	**0.714**
	Baseline	0.723	51.353	0.135	0	0	0	0	0.678

DCI-UG$_b$ in terms of MAE. However, the contrary happens for accuracy, where *DCI-UG$_b$* is slightly better than *DCI-UG$_a$*. Regarding the differences in terms of F-measure per class, we observed that both submissions were very competitive for *pol-5* obtaining even better rates than the best ranked system. Indeed, for this particular class *DCI-UG$_b$* ranked as the second place with a difference of 0.003 with respect to the highest outcome. For what concerns to the underrepresented classes in the training set, our proposals had a lower performance than the best ranked systems. It is interesting to note that, *DCI-UG$_a$* was not able to correctly classify instances of *pol-1* while just a few of *pol-2*, the contrary occurs for *DCI-UG$_b$* with the difference that it ranks at the fourth position for *pol-1*.

3 Conclusions

In this paper, we described our participation for Sentiment Analysis over touristic comments in the framework of the REST-MEX shared task. The proposed approach is based on a modification of the well-known BERT architecture. A Spanish BERT-base model was used. The training data provided by the task' organizers was enriched by using additional data, and during the development stage a subset of the official test partition was manually annotated for validation purposes. Our systems ranked at the 4^{th} and 6^{th} positions according to the official results. As future work, we are interested in to enrich our model considering other type of information (such as for example related to affect and psycholinguistics) and also to deal with the presence of figurative language devices such irony and sarcasm.

References

1. Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. *Sentiment Analysis in Social Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2016.
2. Bharat Gaind, Varun Syal, and Sneha Padgalwar. Emotion Detection and Analysis on Social Media. *CoRR*, abs/1901.08458, 2019.

3. Saif M. Mohammad. Challenges in Sentiment Analysis. In *A Practical Guide to Sentiment Analysis*. Springer, 2016.
4. Delia Irazú Hernández Farías and Paolo Rosso. Irony, Sarcasm, and Sentiment Analysis. Chapter 7. In *Sentiment Analysis in Social Networks*, pages 113–127. Morgan Kaufmann, 2016.
5. Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, and Elio A. Villaseñor. A Case Study of Spanish Text Transformations for Twitter Sentiment Analysis. *Expert Systems with Applications*, 81:457–471, 2017.
6. Isabel Segura-Bedmar, Antonio Quirós, and Paloma Martínez. Exploring Convolutional Neural Networks for Sentiment Analysis of Spanish Tweets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1014–1022, 2017.
7. Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, 9(3), 2020.
8. Daniel Palomino and José Ochoa-Luna. Advanced Transfer Learning Approach for Improving Spanish Sentiment Analysis. In *Advances in Soft Computing*, pages 112–123, Cham, 2019. Springer International Publishing.
9. Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. Learning Sentiment Lexicons in Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3077–3081. ELRA.
10. Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54, 2009.
11. Carlos Henriquez Miranda and Jaime Guzman. A Review of Sentiment Analysis in Spanish. *TECCIENCIA*, 12:35–48, 12 2016.
12. María Navas-Loro and Víctor Rodríguez-Doncel. Spanish Corpora for Sentiment Analysis: A Survey. *Language Resources and Evaluation*, 54, 06 2020.
13. Miguel Á Álvarez-Carmona, Ramón Aranda, Samuel Arce-Cárdenas, Daniel Fajardo-Delgado, Rafael Guerrero-Rodríguez, A. Pastor López-Monroy, Juan Martínez-Miranda, Humberto Pérez-Espinosa, and Ansel Rodríguez-González. Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism. volume 67, 2021.
14. Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *CoRR*, abs/1808.01974, 2018.
15. José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.
16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *CoRR*, abs/1706.03762, 2017.
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018.
18. Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Swish: A Self-Gated Activation Function. *arXiv: Neural and Evolutionary Computing*, 2017.
19. Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 12 2014.
20. Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. Reducing BERT Pre-Training Time from 3 Days to 76 Minutes. *CoRR*, abs/1904.00962, 2019.