Cascade of Biased Two-class Classifiers for Multi-class Sentiment Analysis.

José Abreu[1][0000−0002−4637−4206], Pedro Mirabal[2][0000−0001−7345−6007], and Adrián Ballester-Espinosa[3][0000−0003−2506−1785]

[1] U.I. for Computer Research. University of Alicante. Spain.
ji.abreu@ua.es
[2] Departamento de Ingeniería Informática. Unversidad Católica de Temuco. Chile.
pedro.sanchez@uct.cl
[3] Department of Software and Computing Systems. University of Alicante. Spain.
adrian.ballester@ua.es

Abstract. In this paper, we describe our participation in the Rest-Mex 2021 Sentiment Analysis Task. Our approach is based on an ensemble of BERT|BETO-based classifiers arranged in a cascade of binary models trained with a bias towards specific classes with the aim of lowering the Mean Average Error. The resulting models were judged in the 2nd and the 3rd place according to the evaluation rule of the Mean Absolute Error.

Keywords: Sentiment Analysis · Deep Learning · Transformer Models.

1 Introduction

Sentiment Analysis is a branch within Natural Language Processing that helps us to analyze the opinion of people as regards different entities such as services and products, classifying them into different categories. It is possible to consider positive, negative, or neutral classes or other more fine-grained scales. This task has received notable attention since stakeholders can leverage data from social media or specialized websites like Tripadvisor to make data-driven decisions. However, there are challenges, for example, the uneven development of resources for different languages [1].

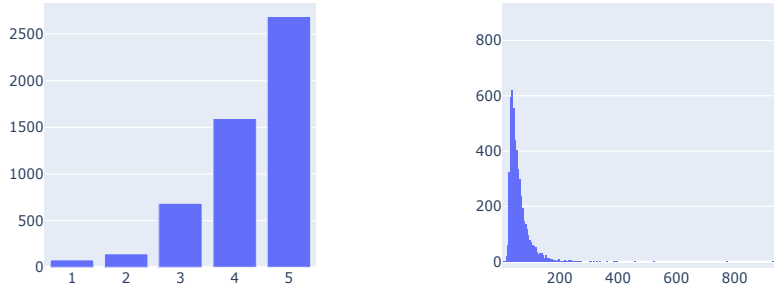
To promote Sentiment Analysis, several challenges have been created on this subject such as SemEval starting in the 2007 edition, IberLEF, and lately Rest-Mex. Recently, Sentiment Analysis has been enhanced by Deep Learning, the details of this topic can be seen in a survey titled Deep learning for sentiment analysis: A survey[2]. Using similar strategies, other teams have participated in competitions related to this field, obtaining good results [3,4,5].

In this paper, we describe our participation in Rest-Mex 2021 Sentiment Analysis Subtask [6]. This subtask is a classification task, the objective is that

IberLEF 2021, September 2021, Málaga, Spain.

systems can predict the polarity of an opinion, published by a tourist about places of Guanajuato, Mexico. The collection provided was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2020. Our approach is based on Deep Learning Transformer Models, specifically BERT and BETO, applying a particular architecture and training strategies that we describe in the next section. The resulting models were judged in 2nd and 3rd places according to the evaluation rule of the Mean Absolute Error.

2 Task and Data Description



(a) Class distribution.

(b) Frequency of opinion length (after BERT tokenizer)

Fig. 1. Training data statistics.

In this section, we describe the data provided by the organizers for this subtask and its characterization. The corpus consists of 7.632 opinions shared, where 5.784 opinions are from national tourists (from Mexico) and 1.848 opinions come from Iberoamerican tourists and the different results of our models. 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 nationality and gender. The organizers split the corpus $70\% - 30\%$ approximately. 70% of the data was delivered to the participants with complete information about each opinion, specifically 5194 opinions. 30% was reserved for the final testing of competing models. 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 $2,688$ instances, representing 51.75% of the

total, a great contrast with the class 1, for which only 80 instances were provided, for the 1.54%. The presence of the rest of the classes is as follows: 1595 instances for class 4, 686 instances for class 2 and 155 instances for class 2, each representing 30.71%, 13.21% and 2.98% respectively. This information can be viewed in Fig. 1a. Our work takes as primary data, only the textual information of the opinion, without taking into consideration other features. One aspect to take into account given the architecture used is the length of each opinion since our model is limited to 512 tokens. In Fig.1b, we show a histogram, where it can be seen that the opinions processed meet this condition.

3 System architecture.

In this section we describe the two architectures, shown Fig.2, we explored for the sentiment analysis task.

3.1 BERT—BETO-based multi-class classifiers.

This model is a multi-class classifier learning the five categories simultaneously. It is based on BERT [9] as feature extractor, fine-tuned for the text classification downstream task. We evaluated two versions of the architecture depicted in Fig. 2a. In both cases, we leveraged transfer learning from pre-trained embeddings. The first one is the uncased version of BETO[4] [8], which is a BERT model trained on a spanish language corpus. The other pre-trained embedding we use is the multilingual uncased version of BERT[5]. We aim to compare a model specific for Spanish to a multilingual one.

The classifier comprised a dense layer with 768 hidden units and RELU activation. Dropout with a rate of 0.2 and a dense layer with 5 units and linear activation. For both BETO and BERT we use the base version, i.e. token embedding of size 768 and max length of 512. As shows Fig. 2a the embedding of the [CLS] token was used as the representation for the whole opinion. To address the unbalanced problem, class weight was set proportional to the number of instances in each category.

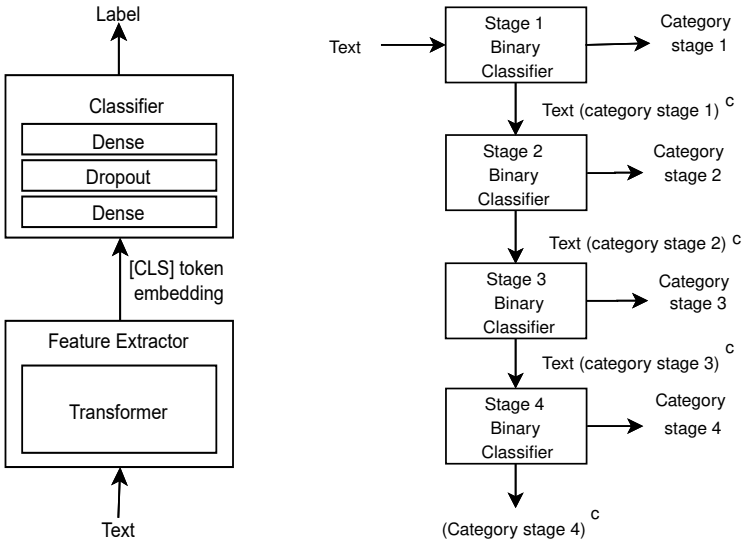
This architecture has been evaluated by the authors of BERT [9] for the sentiment analysis task over the Stanford Sentiment Treebank dataset [10] achieving state-of-the-art results at the time. This makes the architecture attractive as a benchmark for the sentiment analysis task in Spanish.

3.2 BERT—BETO-based two-class cascade classifiers.

The other model we studied is an ensemble of binary classifiers arranged in cascade, as shown in Fig. 2b. Cascading classifiers is the strategy leveraged

[4] `https://github.com/dccuchile/beto`
Available through HuggingFace library, model id: 'dccuchile/bert-base-spanish-wwm-uncased'

[5] `https://github.com/google-research/bert/blob/master/multilingual.md`
Available through HuggingFace library, model id: 'bert-base-multilingual-uncased'



(a) BERT-based multi-class.

(b) Cascade classifier.

Fig. 2. Architectures of the two models studied.

by widely known frameworks such as the Viola-Jones one [11]. In Sentiment Analysis, it has been used by [7] to enrich the feature set of a classifier.

We evaluated two different ways to use this architecture to solve the five-category classification problem. Let's denote the target category at stage i as C_i for the sake of conciseness. The first setup teach each classifier to tear apart instances from one class from the rest. For the classifier at stage 1, $C_1 = \{1\}$ while $C_1^c = \{2, 3, 4, 5\}$. The model at stage 2 learn to classify $C_2 = \{2\}$ and $C_2^c = \{1, 3, 4, 5\}$ and similarly for stage 3. For the last stage, we set $C_4 = \{5\}$ and $C_4^c = \{1, 2, 3, 4\}$.

The other setup explored biasing the classifiers so they tend to classify as the stage target category the miss-classified instances from upstream classifiers. In this case for stage 1 we have $C_1 = \{1\}$ and $C_1^c = \{2, 3, 4, 5\}$. For stage 2, $C_2 = \{1, 2\}$ and $C_2^c = \{3, 4, 5\}$. Note that in this case, stage 2 will also consider category 1 as the target class. We proceeded analogously for the other stages except for the last one which is configured as $C_4 = \{5\}$ and $C_4 = \{1, 2, 3, 4\}^c$.

To classify an instance, we present it to the classifier at stage 1. If classified as the stage target category, then we are done. Else, i.e classified as C_1^c, the instance is presented to the next step classifier. This process is repeated for each stage to the end.

Each classifier is a binary version of the model described in section 3.1 all of them trained separately. For each of these models, we evaluated the multi-language BERT [9] and BETO [8], yielding four different approaches.

4 Results

After the data was processed, it was divided into $90\% - 10\%$ for training and validation. During the fine-tuning process, special attention was paid to different evaluation criteria such as MAE and balanced accuracy.

As a result of the experimentation, 4 new models were obtained. In Table 1, we show in total 6 models, because we include benchmark models for BERT and BETO respectively. Following, we will describe each of these 6 models. Table 1 shows models in descending order, based on the MAE values. At the bottom, we can see the BERT Multi model, which is a BERT model trained on a multilingual corpus, wet considered that model as our Benchmark. The rest of the labels of each model can be interpreted using the following nomenclature: Multi indicates that the model was trained with a multilingual corpus. Biased or Unbiased, refers to what has been stated in Subsection 3.2.

The BETO Biased model was the one that obtained the best result considering its MAE value of 0.51, so it was selected as our primary submission, along with it, the BETO Multi model was sent as secondary submission, with MAE of 0.53. With the selection of BETO Multi model as a secondary submission, we wanted to validate the two cascade variants proposed in this paper. As the final stage before submitting, the two selected models were trained with the total available data.

	MAE		RMSE		Acc.		F1		Rec.		Prec.	
	train	val	train	val	train	val	train	val	train	val	train	val
BETO Biased (sub1)	0.03	0.51	0.06	0.65	0.95	0.44	0.92	0.40	0.95	0.44	0.90	0.40
BETO Unbiased	0.04	0.53	0.08	0.68	0.95	0.48	0.90	0.44	0.95	0.48	0.87	0.49
BERT Multi Biased	0.08	0.53	0.13	0.68	0.94	0.45	0.87	0.43	0.94	0.45	0.84	0.48
BERT Multi Unbiased	0.26	0.67	0.71	1.23	0.80	0.47	0.70	0.36	0.80	0.47	0.74	0.38
BETO Multi (sub2)	0.39	0.53	0.50	0.73	0.74	0.53	0.65	0.48	0.74	0.53	0.62	0.48
BERT Multi	0.62	0.70	0.91	1.05	0.59	0.51	0.45	0.40	0.59	0.51	0.41	0.37

Table 1. Experiment Results.

In this subtask, 8 teams competed, with 14 submissions in total. In the team ranking, we were second, and in the submission ranking we were in second and third place, our best-evaluated submission turned out to be the secondary one, with 0.5451 of MAE. It should be noted that our best submission achieved the best result in F-measure and Precision among all participants. A summary of the competition can be seen in Table 2.

5 Conclusion and Future Work

In this paper, we have described the models proposed by UCT-UA in the Sentiment Analysis subtask at Rest-Mex 2021. We presented two models, the results

Team Rank	Sub. Rank	Team	MAE	RMSE	Accuracy	F-measure	Recall	Precision
1st	1	Minería UNAM 1	**0.4752**	**0.7549**	**56.7238**	0.4280	**0.4992**	0.4035
2nd	2	**UCT-UA 2**	0.5451	0.8540	53.2491	**0.4512**	0.4662	**0.4933**
	3	**UCT-UA 1**	0.5614	0.9023	53.8357	0.4035	0.3984	0.4626
3rd	4	DCI-UG 1	0.56273	0.8843	53.3394	0.2870	0.3405	0.2827
	5	Minería UNAM 2	0.5826	0.9498	54.7834	0.2428	0.2732	0.2549
	6	DCI-UG 2	0.6060	0.97046	53.7004	0.2539	0.3004	0.2772
		BASELINE	0.7238	1.1620	51.3538	0.1357	0.1027	0.200

Table 2. Competition Ranking.

in our secondary submission were obtained from the model described in the Results section as BETO Multi, this model the second-best result in the subtask achieving 0.5451 of MAE. The results in our primary submission were obtained from the model described in the Results section as BETO Biased, this model the third-best result in the subtask achieving 0.5613 of MAE.

Comparing to the models using BERT Multi, the results suggest that the monolingual embedding is a better representation. However, this is consistent with results from BERT team [6]where for high-resource languages the multilingual model may achieve the worst results respect the single-language model. Moreover, this can be aggravated since the fine-tuning was done using Spanish only thus degrading the multilingual representation spaces spawned by the transformer.

As future work, we are interested in evaluating if the multilingual models can benefit from Tripadvisor reviews in different languages or topics. Also, we would like to study multi-modal approaches that can leverage information from the title or metadata of the review to boost the results.

6 Acknowledgments

This research work has been partially funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) and the Spanish Government through the projects SIIA (PROMETEO/2018/089, PROMETEU/2018/089) and LIVING-LANG (RTI2018-094653-B-C22), and the Vice Chancellor for Research and Postgraduate Studies Office of the Universidad Católica de Temuco, VIPUCT Project No. 2020EM-PS-08; FEQUIP 2019-INRN-03 of the Universidad Católica de Temuco.

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)

[6] https://github.com/google-research/bert/blob/master/multilingual.md

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. 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)
7. 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)
8. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: Proc. of PML4DC at ICLR. (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. (2018) https://doi.org/arXiv:1810.04805
10. 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)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In. Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. (2001)