Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor

Juan Vásquez[1], Helena Gómez-Adorno[2], and Gemma Bel-Enguix[1]

[1] Instituto de Ingeniería,
Universidad Nacional Autónoma de México, Mexico City, Mexico
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México, Mexico City, Mexico
juanmvs@pm.me, helena.gomez@iimas.unam.mx, gbele@iingen.unam.mx

Abstract. This paper presents our approach to the Sentiment Analysis Task at REST-MEX 2021. The goal of the task is to predict the polarity of opinions on Mexican tourist sites. Recent advances in transfer-learning with pre-trained models in English have produced state-of-the-art results in sentiment analysis. In this work, we apply two Bert-based approaches for review classification in five classes. Our first approach consists of fine-tuning Beto, a Bert-like model pre-trained in Spanish. Our second approach focuses on combining Bert embeddings with the feature vectors weighted with TF-IDF. The results obtained using the standalone BERT model ranked first in the task.

Keywords: Sentiment analysis · Opinion mining · Spanish language · BERT · Transfer learning

1 Introduction

Sentiment analysis is an area of research in Natural Language Processing (NLP) whose goal is to extract a sentiment or emotion of a given opinion [3]. It is considered a classification problem [7] where one sentiment is assigned to an opinion. Recent advances in transfer learning have greatly improved the state of the art in many NLP problems. Bert, for example, has achieved accuracy as high as 89.7 in binary classification [4].

To encourage the research community in NLP to develop new research areas in Spanish, the Iberian Language Evaluation Forum (IberLEF) organizes yearly a comparative evaluation campaign [5]. In 2021, the REST-MEX task [1] (Recommendation System for Text Mexican Tourism) was proposed, including two sub-tasks. The first one focused on a recommendation system for tourist sites given a tourist's profile and their personal preferences. The second task required a sentiment analysis system able to classify an opinion about a Mexican tourist

IberLEF 2021, September 2021, Málaga, Spain.

place with a score between 1 and 5. Our team focused only on the second task. It is worth noticing that our first submitted run obtained the best results in the second subtask.

This paper is organized as follows: Section 2 states the task we worked on and the data sets provided by the task organizers; Section 3 describes the two systems that we designed and implemented; Section 4 details our experiments and the obtained results; and Section 5 reports our conclusions.

2 Task and Data Description

The sentiment analysis task required the prediction of a class for each review provided in the evaluation set. The available classes were integers between 1 and 5. The reviews were taken from the website TripAdvisor and were written by a tourist who evaluated a landmark in Guanajuato, Mexico. All of them were in Spanish.

The participants were provided with two different data sets; one for training and one for evaluation. The training set was made up of 5197 rows and 9 columns described as follows:

1. Index: The index of each opinion.
2. Title: The title that the tourist himself gave to his opinion.
3. Opinion: The opinion expressed by the tourist.
4. Place: The tourist place that the tourist visited and to which the opinion is directed.
5. Gender: The gender of the tourist.
6. Age: The age of the tourist at the time of issuing the opinion.
7. Country: The Country of origin of the tourist.
8. Date: The date the opinion was issued.
9. Label: The label that represents the polarity of the opinion: [1, 2, 3, 4, 5].

For our experiments, we only trained the models with the "Opinion" column.

The classes in the training set were not balanced. The distribution can be seen in Table 1

Table 1. Labels Distribution

Label	Counts	Distribution (%)
5	2688	51.8
4	1595	30.7
3	686	13.2
2	145	2.8
1	80	1.5

The evaluation set contained 2216 rows and the same first 8 columns listed for the training set.

3 Proposed Approaches

At 2020's edition of the task on Semantic Analysis at SEPLN, a Bert-like model for sentiment analysis at three levels yielded the highest accuracy [6]. The second best results were also obtained by applying a system based on Bert [9]. Even though these architectures were designed to classify tweets, their results motivated us to work on Bert-based approaches.

Another reason for working with Bert was that fine-tuning it for downstream tasks, such as sentiment analysis, is computationally inexpensive [4]. This process for sentiment analysis is made up of feeding Bert with task-specific inputs and passing the outputs through a classification layer. The results obtained in the original Bert paper established a new state of the art in sentiment analysis at three levels [4].

Because the reviews in the data sets were written in Spanish, we decided to use Beto as our baseline model. This Bert-like system was pre-trained using a corpus in Spanish with a similar size to that of the corpus used to train Bert-Base [2]. Considering that Beto follows the same design principles as Bert, we proceeded to execute a fine-tuning process for five classes.

Section 3.1 describes the approach for fine-tuning Beto, which ranked first in this Task. Section 3.2 outlines a second approach we took in hopes of improving our results. In this system, we added a Bag-of-words feature vectors weighted with TF-IDF, to the contextual embeddings generated by Beto after fine-tuning it. The motivation behind this was that TF-IDF captures global information from all the entries in the data set [8], while Bert only captures contextual information from the attention mechanism [4].

It is important to mention that, before proceeding with the fine-tuning, we performed a data pre-processing step, which consisted of removing the quotation marks in the reviews.

The code for implementing the systems listed here can be found in our Github repository[3].

3.1 Fine-tuned Bert approach

First, we passed the reviews through the Bert tokenizer (loaded with the weights from Beto). Once the tokens were generated, we fed them to Beto with a classification layer on top. This step executed the domain-specific training (or fine-tuning). The next step was to repeat this same process on the official evaluation set. Once we tokenized the reviews in this second set, we predicted the classes by passing these tokens through the previously fine-tuned model.

The hyperparameters used for the fine-tuning were the ones recommended in the original Bert paper [4]. These were:

- Max length = 512
- Batch size = 8

[3] https://github.com/juanmvsa/Sentiment-Analysis-TripAdvisor-Spanish

- Optimizer = AdamW
- Learning rate = $2e - 5$
- Steps = $1e - 8$
- Epochs = 4

3.2 Fine-tuned Bert Approach with TF-IDF vectors

We started by extracting the contextual embeddings generated after fine-tuning Beto with the training set (we followed the same steps listed on Section 3.1). Then, we obtained a new set of features by first tokenizing the original training set using Spacy's model "es_dep_news_trf". Next, we calculated the TF-IDF weights for those tokens. Once we had the two set of features, we concatenated them into one set, which we then used for training a logistic regression algorithm. Next, we generated the corresponding contextual embeddings from Beto and TF-IDF bag-of-words features for the evaluation set. Finally, we predicted the classes with the previously trained logistic regression model.

For this system we utilized the same hyperparameters listed on 3.1.

4 Results

For this competition, the ranking was determined by measuring the systems with the mean absolute error (MAE). As can be seen in Equation 1, this metric outputs a final number which is calculated by summing the magnitudes (absolute values) of the errors to obtain the "total error" and then dividing the total error by n [10].

$$MAE = n^{-1} \sum_{i=1}^{n} |e_i| \tag{1}$$

In order to get an overview of how our systems performed, we tested various classification algorithms on a partition of the training set. For all the experiments in Table 2, we tokenized the reviews using the Spacy model "es_dep_news_trf". The next step was to generate the TF-IDF weights. Then, we trained the different classification algorithms. Finally, we evaluated each model.

The results in Table 2 show the performance of the supervised algorithms, of the fine-tuned Beto, and of Beto with the added TF-IDF feature vectors. It is observed that fine-tuning Beto yields the best results in this task.

Table 3 presents the results provided by the organizers of the competition. These metrics were obtained on the evaluation set.

5 Conclusions

The approach described in 3.1 obtained the lowest mean absolute error in the Subtask 2 of this year's sentiment analysis shared task at IberLef. This suggests that transfer learning is a functional approach for sentiment analysis with five

Table 2. Obtained results when training on 80% of the training set and testing on the 20% left

Approach	**MAE**
BOW + TF-IDF + Logistic regression	0.7632
BOW + TF-IDF + SVM	0.5592
BOW + TF-IDF + Linear regression	0.8499
BOW + TF-IDF + Lasso regression	0.7709
BOW + TF-IDF + Elastic net regression	0.7709
BETO FT (simple)	**0.4519**
Beto FT + TF-IDF + Logistic regression	0.5850

Table 3. Obtained results on the evaluation set

Approach	**MAE**	**RMSE**	**Accuracy**	**F1-score**	**Recall**	**Precision**
Run 1: Beto FT	**0.4751**	0.7549	56.7238	0.4280	0.4992	0.4981
Run 2: Beto TF-IDF	0.5825	0.9497	54.7833	0.2428	0.2731	0.2548

classes. Even though we achieved the lowest mean absolute error among all the teams in the shared task, we consider that this number is still very high. This could be due to the difficulty of doing classification among five classes.

One limitation of our approach is that transfer learning depends on the corpora used to pre-train the model. This restricts the learning capabilities to certain topics. Also, Beto was pre-trained using a corpus with a size similar to that of Bert-Base. We hypothesize that using a pre-trained model with more parameters would greatly improve our results.

Another limitation of Beto is the computational power. Until now, these architectures can only deal with a maximum length of 512 tokens per sequence. This means that some reviews that exceed that token length are truncated before the encoding, leading to a loss in data, and, therefore, in learning.

Furthermore, we propose working on different approaches to multi-class classification. While our second system did not produce better results than the first one, it did achieve the fifth place among the participants. We hypothesize that combining different sets of features generated by different language models could improve the classification results when faced with five classes.

Finally, by manually analyzing the data set, we observed that sometimes the classes had very little relation to the review. For example, the following review was labeled with class 1: “If you go as a couple this place is a must, it is special to climb and kiss on the third step, very romantic and emblematic”. When reading the review, one can not infer that this would be given such a low class. This kind of annotation in the data set could be the reason behind the poor performance obtained in this task.

Acknowledgments

This research was partially funded by CONACyT project CB A1-S-27780 and DGAPA-UNAM PAPIIT grants number TA100520 and TA400121.

References

1. Álvarez-Carmona, M.Á., Aranda, R., Arce-Cárdenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A.P., Martínez-Miranda, J., Pérez-Espinosa, H., Rodríguez-González, A.: Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism. Procesamiento del Lenguaje Natural **67** (2021)
2. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
3. Dale, R., Somers, H.L., Moisl, H.: Handbook of Natural Language Processing. Marcel Dekker, Inc., USA (2000)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Forum, I.L.E.: Iberian languages evaluation forum (February 2021), `https://sites.google.com/view/iberlef2021`
6. González, J., Moncho, J.A., Hurtado, L., Pla, F.: Elirf-upv at tass 2020: Twilbert for sentiment analysis and emotion detection in spanish tweets. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain. vol. 23 (2020)
7. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal **5**(4), 1093–1113 (2014)
8. Meng Lim, W., Tayyar Madabushi, H.: Uob at semeval-2020 task 12: Boosting bert with corpus level information. arXiv e-prints pp. arXiv–2008 (2020)
9. Palomino, D., Ochoa-Luna, J.: Palomino-Ochoa at tass 2020: Transformer-based data augmentation for overcoming few-shot learning. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) (2020)
10. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research **30**(1), 79–82 (2005)