UC3M-sas at IberLEF2024 DETESTS-Dis tasks

Adrián Gónzalez Sánchez[1,†], David Santiago García-Chicangana[1,†] and Santiago Rondón Galvis[1,†]

[1]Carlos III University of Madrid

Abstract

This paper presents our work made for the DETESTS-Dis IberLEF 2024 competition, aimed at detecting racism stereotypes in spanish texts. The competition task comprises two subtasks: the detection of stereotype in texts, and to identify wheter a stereotype is explicitly or implicitly stated in the text. In this paper, we detailed the data used and the approach implemented as well as the challenges found throughout the process. As a result, we obtained an F1-score of 0.641 and a cross entropy of 0.841 with hard and soft labels respectively.

Keywords

Transformers, BERT, DETESTS, racial-stereotypes

1. Introduction

A stereotype, as defined by social psychology, is a set of beliefs about individuals who are perceived to belong to a different social category. These stereotypes simplify the group by generalizing a single characteristic and applying it to all members of the group [1]. One way in which stereotypes manifest themselves is through language. They can be in the form of explicit or implicit expressions, making them a complex concept when trying to operationalize them for natural language processing.

In natural language processing (NLP), several efforts have been addressing this problem, including initiatives such as IberLEF that seek to foster research in this area [2]. In this article, we present our work for the IberLEF 2024 competition, specifically in the task for detecting and classifying racial stereotypes in Spanish texts (DETESTS-Dis) [3]. This task comprises two binary classifications: The first task is aimed at detecting whether a stereotype is present or not in the text, while the second consists of detecting if that content is explicitly or implicitly stated in the text. To achieve this, transformer-based models trained with a large Spanish corpus were employed for this task. The evaluation metrics used for the first task were the F1-score for hard labels and cross entropy for soft labels. In the case of the second task, the ICM metric was applied to both types of labels.

The structure of the paper is as follows: Section 2 and Section 3 present a description of the dataset provided for the challenge and the data augmentation techniques used to handle the imbalanced dataset, respectively. Section 4 describes the transformer-based models used in this

IberLEF 2024, September 2024, Valladolid, Spain

[†]All authors contributed equally.

100494633@alumnos.uc3m.es (A. G. Sánchez); 100508815@alumnos.uc3m.es (D. S. García-Chicangana); 100506421@alumnos.uc3m.es (S. R. Galvis)

work, and Section 5 presents the experiments conducted and the results obtained throughout this process. Finally, Section 6 provides the conclusions of this work.

2. Dataset

The provided dataset contains 9906 records, which were collected from two main sources: 5629 records extracted from comments on news articles (DETEST), and 4722 from Tweets of 2021 reacting to hoaxes. In total, 7301 records were tagged as non-stereotypes and 2605 with stereotypes. The label consists of a value between 0 and 1, where the number 1 indicates the presence of stereotypes in the text, and zero its absence. The records tagged with stereotypes represent 26.29% of the total records, which clearly shows an imbalance in the dataset that can affect the results of our models.

For the second task, from the 2605 records tagged with stereotypes, 1279 were marked as explicit and 1329 as implicit. Although the size of the records with stereotypes is small, it presents a better balance between the explicit and implicit stereotypes stated in the text records.

Both datasets were tagged using hard labels with values of 0 and 1 that represent the vote of the majority of the dataset evaluators. Soft labels, with values between 0 and 1, were also included in the dataset. These values result from the softmax function applied to the evaluator scores.

For the training phase of the models, we divided the initial dataset in the following way: 70% for training, 15% for validation, and 15% for testing. In this way, we evaluate the proposed models to see which obtain the best result. Once the testing dataset was released, we changed the distribution of the initial dataset, being 80% for training, and 20% for validation.

3. Data Augmentation

Due to the highly imbalanced distribution of classes for task 1, some augmentation techniques were applied, not in the whole dataset, but in the training set. The first technique was Back-Translation [4], using the Python Translate library to translate the entire initial dataset into English. Then, we translated it back into Spanish. The model (*Helsinki-NLP/opus-mt-es-en*[1]) loads a translation model from Spanish to English, and (*Helsinki-NLP/opus-mt-en-es*[2]) loads the opposite model, from English to Spanish. The second technique, Synonym Substitution [5] [6], allows us to replace words in the original text with their synonyms, thus generating variants of the same text but keeping its original meaning. It helped us to improve the robustness of the dataset to obtain better results when training the model. To perform this process, the 'nlPaug' library was used with a variation of the Bert model in Spanish, such as 'dccuchile/bert-base-spanish-wwm-uncased'[3] [7], and iterating over all the rows of the training set to generate the new versions of each text.

The results obtained from the application of these techniques are presented in Table 1.

[1]https://huggingface.co/Helsinki-NLP/opus-mt-en-es
[2]https://huggingface.co/Helsinki-NLP/opus-mt-es-en
[3]https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

Task	Back-Translation	Synonyms substitution	total training set
Task 1 - Hard labels	2319	1437	11680
Task 1 - Soft labels	(2319+286*+1031**)	0	11560

Table 1
Result from the application of the Data augmentation techniques. * Back-Translation applied to softlabels with a value between 0.5 and 0.9. ** Back-Translation applied to softlabels with a value between 0.1 and 0.5

The Back-Translation technique was applied to the texts of the training set with label 1, resulting in 2319 new texts. In the case of the second technique, 1437 new records were generated. All these records helped us to achieve a balance between both classes, thus obtaining a total of 11680 records, 5840 with label 1 and 5840 with label 0.

Regarding Task 1 with soft labels, we found that 6270 records of the training set contain the value of 0.0474 for the column *stereotype_soft*, 2319 records with the value 0.9526, 1031 records with the value 0.2689, and 286 with a value of 0.7311. As we can see, the last two values contain fewer instances. To minimize this imbalance, Back-Translation was applied to these records, generating 286 and 1031 new records for *stereotype_soft* values of 0.2689 and 0.7311, respectively.

No augmented data was used for the second task due to the balance between explicit and implicit stereotypes present in the dataset.

4. Models

One of the approaches applied was using transformer models to take advantage of the latest advancements in the field of NLP. The models used in the work are presented below.

In general, the models were loaded using code libraries such as *pip*, *transformers*, and *accelerate*. The training was executed in Google Colab using Tesla T4 GPU machines. The datasets provided by the organizers were in CSV format and loaded using the Pandas library.

For the evaluation metrics, we used sklearn's `classification_report`[4], which provides a detailed description of the precision, recall and *F1-score* metrics for each *implicits* and *explicits* class. This is because these metrics provide a balanced evaluation of the model, especially for unbalanced datasets, as was the case for the given dataset initially. With the `classification_report`, the precision indicates the accuracy of the positive predictions, the recall measures the ability of the model to correctly identify positive instances and finally, the *F1-Score* value is the one that gives an average between these two aspects to have a metric that reflects the precision and recall. In addition, we used a confusion matrix to have a clear visualization of the model's errors, which allowed us to identify patterns of errors or areas for improvement.

4.1. BERT

For the first task, we decided to fine-tune the model BERT, which translates to Bidirectional Encoder Representations from Transformers. This model, developed by Google in 2018 [8], is

[4]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

particularly popular given its high accuracy in natural language processing tasks when initially presented. It was presented as a technological innovation since, unlike other models used at the time, it was the only one of its kind. At that time, most models could only be trained in a specific way, so when BERT arrived with bidirectional training, there was an understanding of the flow and context of language, compared to other models that could only train in one direction. In addition, this model in its original version was trained with a huge dataset, such as the 'Toronto Book Corpus'[5] and all data from Wikipedia. Precisely, this was one of the main reasons why we selected BERT to perform both tasks. Since it has bidirectional text comprehension, the model can consider the context before and after the word, giving us a much more accurate understanding and meaning according to its context. In addition, since this model is pre-trained with a large dataset, as mentioned above, it allows the model to understand a variety of linguistic and cultural nuances, making it very useful for recognizing subtle patterns in Spanish, including, of course, stereotypes, giving us a more accurate classification. Finally, it should be noted that although the BERT model was primarily trained in English, there are also variants in other languages, such as the BETO model [7], which was trained specifically for the Spanish language.

4.2. RoBERTa

The RoBERTa (Robustly optimized BERT approach) is a model proposed in [9]. It is based on the BERT architecture and implements a series of improvements in the pretraining procedure to achieve better end-task performance.

Nowadays, we can find some models fine-tuned for different domains. One of the versions, called 'roberta-base-bne'[6], was trained with data from the National Library of Spain (Biblioteca Nacional de España, BNE). Although the authors suggest using it for the fill mask task, they also mention it can be used for other tasks such as text classification, among others. For that reason, we decided to use it for the classification tasks of the challenge, especially since it was trained with one of the largest Spanish corpora (SNE). This model constituted the initial point to start fine-tuning with the data provided to the contestants.

The model had a classification layer where its main output was two values: '0,' which represents the absence of the feature, and '1,' its presence. In the case of the first task, 1 represents the text that contains stereotypes while 0 does not. Regarding the second task, the values 0 and 1 represent the stereotype as explicit or implicit, respectively.

5. Experiments and Results

In the context of NLP models, it is crucial to consider task-specific evaluation metrics. In our research, we have addressed two distinct tasks, each with a different evaluation metric. For the first task, we employed the F1 metric for hard labels and cross-entropy [10] for soft labels. In the second task, the metrics used are the ICM [11], in particular, ICM and ICM Norm for

[5] https://huggingface.co/datasets/bookcorpus
[6] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

hard labels, and ICM Soft and ICM Soft for soft labels. The following paragraphs provide a brief description of the aforementioned metrics.

- **F1**: The F1 score is a harmonic mean of precision and recall. Precision is the proportion of correct predictions among all positive predictions, while recall is the proportion of true positives among all positive instances.
- **Cross Entropy**: This metric measures the discrepancy between two probability distributions: the true distribution (i.e., the actual labels) and the distribution predicted by the model.
- **ICM**: A metric based on information theory that considers both hierarchical class structure and class specificity.
- **ICM norm**: The ICM norm is a normalized version of the ICM metric that adjusts values for easier comparison.
- **ICM Soft**: A variant of the ICM metric that is applied to soft labels, where the model predictions are probabilities instead of discrete labels.
- **ICM Soft norm**: The ICM Soft norm is a normalized version of the ICM Soft metric.

5.1. Experiments

This section will provide a concise overview of the experiments conducted for each task.

Experiment Task 1

For the first task, a set of machine learning models was tested, including CNN, SVM + TF-IDF, and BI-LSTM. The convolutional neural network (CNN) is a type of artificial neural network that has been demonstrated to be effective in the fields of both computer vision and natural language processing (NLP), where it can be employed to capture local patterns in text [12]. The SVM + TF-IDF classifier is a robust supervised classifier that employs TF-IDF to transform a text corpus into a feature matrix, representing word importance [13]. The BI-LSTM is an extension of the standard LSTM that considers both preceding and following context in a text sequence. [14].

Then, we proceeded to test the pre-trained transformer models. In this context, the BETO and RoBERTa-base-bne models were employed, as previously outlined.

Finally, to conclude the experimental phase, we implemented data augmentation on the dataset and re-run the experiments on the BETO and RoBERTa-base-bne models. This was done with the intention of improving the results obtained by using a more balanced dataset.

Experiment Task 2

As this task utilizes the same data as task 1 and the objective is to perform analogous processing, we elected to conduct a single experiment, namely the RoBERTa-base-bne Model, which yielded optimal results in task 1. Upon observing that the outcomes were favorable when the experiment was repeated in this second task, we opted not to conduct further experiments.

Model	Stereotype	No Stereotype	Accuracy	Macro avg
CNN	0.87	0.58	0.80	0.73
SVM + tf-idf	-	-	0.63	-
BI-LSTM	0.86	0.62	0.80	0.74
BETO	0.90	0.71	0.85	0.80
RoBERTa-base-bne	0.90	0.74	0.85	0.82
BETO + Text Augmentation	0.91	0.78	0.87	0.84
RoBERTa-base-bne + Text Augmentation	0.93	0.80	0.89	0.86

Table 2
Results of task 1 experiments using the F1 metric

Model	Implicits	Explicits	Accuracy	Macro avg
RoBERTa-base-bne	0.70	0.77	0.74	0.74

Table 3
Results of task 2 experiments using the F1 metric

5.2. Results

The evaluation of each task in our competition is conducted in a manner specific to that task. To facilitate comparison of the results of the various experiments conducted, we have elected to utilize the F1 metric of task 1 for hard labels. This is because we consider this task to be the most important and because the F1 metric is relatively straightforward to calculate and comprehend, thereby facilitating the improvement of our models.

The initial focus was on Task 1, with the aforementioned experiments conducted. The resulting outcomes are shown in Table 2. It should be noted that in these experiments, we divided the training dataset provided into 3 subdatasets (train 70%, validation 15%, and test 15%) and carried out the experiments on these. Since the test dataset provided to us did not include the final labels, we did not use it until the final training was completed. Finally, for the final training, we divided the training dataset into two parts, with (80%) for training and (20%) for validation. Table 3 presents the results of the sole experiment conducted for Task 2, in which the F1 metric was employed for evaluation purposes.

The initial results of the first three models (CNN, SVM + TF-IDF, and BI-LSTM) were comparable but did not meet the desired performance criteria. Consequently, we opted to utilize pre-trained models. Our observations indicated that the RoBERTa-base-bne model yielded the most favorable outcomes when data augmentation techniques (Back-Translation and synonym substitution) were employed. Upon analysis of the results, it was determined that the RoBERTa-base-bne model would be most suitable for the identification of stereotypes in these tasks.

To this end, the model was once again trained, this time iterating over each hard and soft label per task to generate the results required by the challenge evaluation phase. As a result, our model obtained a high score for task 1, achieving first place in Soft Labels but eighth place in Hard Labels out of a total of 21 submissions. This placed us at the top of the table.

Regarding task 2, we achieved the fourteenth place in hard labels and fifth in soft labels. The poor results obtained in this second task may be attributed to the small size of the dataset,

which contained only 2,500 balanced instances. Future works can focus on applying other data augmentation techniques and different model approaches for these classification tasks.

6. Conclusions

This article presents our work for the two classification tasks of the DETESTS-Dis challenge[3] belonging to IberLEF2024[7]. The objective was to identify whether sentences contain stereotypes and, if so, to determine whether these are explicit or implicit. The absence of implicit bias necessitates the implementation of two distinct tasks of binary classification. According to our results, the RoBERTa-base-bne model, along with techniques of data augmentation, resulted in an F1-score of 0.641 for the hard labels and a cross-entropy of 0.841 for soft labels, respectively. The latter value is the best for the soft label category in this first task. However, the outcomes of the second task were less impressive, with the team ranking towards the lower end of the table. For future works, other techniques for data augmentation can be applied, such as using generative models like GPT to create new instances and increase the number of records with stereotypes. Besides, new models based on transformer architectures can be useful for testing in classification tasks, such as GPT and Llama, among others.

Acknowledgments

We express our gratitude to Professor Isabel Segura Bedmar for her support and teachings throughout this process. Her guidance and encouragement have been instrumental in the completion of this research.

References

[1] G. W. Allport, The nature of prejudice, Addison-Wesley Pub. Co., Cambridge, Mass, 1955.
[2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
[3] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, M. Taulé, Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learn with Disagreement, Procesamiento del Lenguaje Natural 69 (2024).
[4] A. Sugiyama, N. Yoshinaga, Data augmentation using back-translation for context-aware neural machine translation, in: A. Popescu-Belis, S. Loáiciga, C. Hardmeier, D. Xiong (Eds.), Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 35–44. URL: https://aclanthology.org/D19-6504. doi:10.18653/v1/D19-6504.

[7] https://sites.google.com/view/iberlef-2024/home

[5] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.

[6] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, 2016. arXiv:1509.01626.

[7] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[8] 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.

[9] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[10] P.-T. de Boer, D. Kroese, S. Mannor, R. Rubinstein, A tutorial on the cross-entropy method, Annals of operations research 134 (2005) 19–67. doi:10.1007/s10479-005-5724-z.

[11] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[12] Y. Kim, Convolutional neural networks for sentence classification, 2014. arXiv:1408.5882.

[13] J. Lilleberg, Y. Zhu, Y. Zhang, Support vector machines and word2vec for text classification with semantic features, in: 2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), 2015, pp. 136–140. doi:10.1109/ICCI-CC.2015.7259377.

[14] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, J. W. Kim, Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/17/5841. doi:10.3390/app10175841.