I2C at IberLEF-2022 DETESTS task: Detection of Racist Stereotypes in Spanish Comments using UnderBagging and Transformers

Jacinto Mata Vázquez, Victoria Pachón Álvarez, Chaimae Tayebi Taybi and Pablo Pizarro Sánchez

University of Huelva, Escuela Técnica Superior de Ingeniería, Huelva, Spain

Abstract

This paper presents the systems developed for participation at DETESTS (DETEction and classification of racial STereotypes in Spanish) shared task at IberLEF 2022. The objective of the task is to detect stereotypes in sentences from comments posted in Spanish in response to online news articles related to immigration. The proposed systems are based on pre-trained Spanish language models and the implementation of balancing techniques to tackle the problem of imbalance between comments that contain stereotypes and those that do not. Our systems ranked in the top three positions in subtask 1. An UnderBagging-based classifier reached the first position obtaining an F-score of 0.7042.

Keywords

UnderBagging, Data Augmentation, Language Model, Ensemble, Stereotype, Immigration

1. Introduction

Racism refers to the prejudice, discrimination, or antagonism directed against people of a different race or ethnicity. Today, there are more and more people commenting on articles about immigration because it is an issue of great interest. Thus, social media platforms make it easier for people to express their opinion, but these digital spaces also make it more difficult to avoid people commenting about immigrants. Additionally, it is easy to express opinions anonymously in comments or on social media. The toxicity of immigration-related comments in online spaces has increased significantly in recent years. In this context, one of the first steps to mitigate bias in social media is to detect racial comments in news articles.

In this paper, we present the systems we developed as part of our participation in the DETESTS (DETEction and classification of racial STereotypes in Spanish) shared task [1] at IberLEF 2022 within subtask 1. The aim of the substask is to detect stereotypes in sentences from comments posted in Spanish in response to different online news articles related to immigration.

Due to the recent success of pre-trained language models for addressing document classification tasks, the core of the proposed systems is based on fine-tuning some of the Spanish models to adapt them to this task. On the other hand, due to the significant imbalance between the comments with and without stereotypical messages, different data balancing techniques were used.

The paper is structured as follows. In Section 2, some previous works related to this task are described. Section 3 describes the dataset provided by the organizers and we explain how it is used in the experimentation process. In Section 4, we describe our proposal to address the task and the experiment setup, leading to the results obtained during the practice. Section 5 describes the submitted runs and the findings obtained in the evaluation phase. Finally, we present our conclusions in Section 6.

IberLEF 2022, September 2022, A Coruña, Spain

EMAIL: mata@uhu.es (JMV); vpachon@uhu.es (VPA); chaimae.tayebi820@alu.uhu.es (CTT); pablo.pizarro@alu.uhu.es (PPS)

© 2022 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)

CEUR-WS.org/Vol-3202/detests-paper3.pdf

2. Related works

Some works have been developed on the detection and classification of stereotypes, but mainly based on specific groups, such as women or immigrants. For instance, in [2] the authors propose two experiments for the automatic detection of Dutch language racist speech on social networks. Moreover, in [3] the authors developed and published the first dataset of Spanish language sexist expressions and attitudes on Twitter (*MeTwo*) and investigated the possibility of using machine learning and deep learning techniques for the automatic detection of such stereotypes. The work in [4] is focused on detecting narratives containing xenophobic and conspiratorial stereotypes in articles published on online magazines/web.

A classification task for the identification of racist, sexist and otherwise discriminatory language in online user comments was proposed in [5]. In [6] the authors exploited the implicit knowledge of stereotypes to create an end-to-end stereotype detector using solely a language. In [7], the authors provide a novel proposal to identify stereotypes about immigrants using two different approaches: Transformers and a text masking technique.

3. Datasets and tasks

The training dataset provided by the organizers contains 3817 comments published in response to different articles. There are 18 information attributes available, these are: *comment_id*, *sentence_pos* (position of each sentence within a comment), *reply_to* (identifier of the comment that initiates a thread. It defaults to itself in case it is the first comment in the thread), *sentence* (the comment), *racial_target*, *other_target*, *implicit*, *stereotype*, *xenophobia*, *suffering*, *economic*, *migration*, *culture*, *benefits*, *health*, *security*, *dehumanization*, and *others*. The objective of subtask 1 is to determine if the sentences in a comment contain at least one stereotype or none considering the complete distribution of provided labels.

The dataset contains 2946 comments with the stereotype label at 0, that is, without a racist stereotype message, and 871 comments labeled with 1. On average, 22% of the comments contain racist stereotypes, so the dataset is unbalanced.

Table 1
Some examples of the dataset

comment_id	sentence_pos	reply_to	sentence	stereotype
0	1	0	La solución es desarrollar el pensamiento crítico y escéptico.	0
0	2	0	Hay que enseñar que la magia no existe.	0
5	1	5	Desgraciadamente, y visto lo visto, al final vamos a tener que aplicar lo que hace los israelíes en sus fronteras.	1
1720	2	1720	He visto muchos casos deplorables, propios de ONG´s parapolíticas y bien untadas.	0

We split the provided dataset into 80% for training and validation, and 20% for testing. For the experimentation, this training/validation dataset was divided into 80% to train the models and the remaining 20% for validation. The splits were made in a stratified way to keep the same proportion of class 0 comments and class 1 comments as in the original dataset. Validation dataset was used for hyper-params optimization and early stopping, and test dataset was used for out-of-sample performance estimation. In Table 1 we can see some examples of the dataset and Table 2 shows the number of comments for each class.

Table 2
Class distribution

Dataset	Rows	Class 1	Class 0
Original (provided by the organizers)	3817	871	2946
Train/validation	3053	697	2356
Test	764	174	590

4. Methodology and experiments

Proposed systems for the subtask 1 are based on the fine-tuning of three pre-trained transformer-based models. Moreover, our work is based on using balancing approaches since the provided dataset is unbalanced. These approaches consist of data augmentation [8], undersampling and UnderBagging ensemble [9], which incorporates the strength of random undersampling and the bagging. We also used another ensemble strategy to improve the individual results obtained from each model.

Since the comments are in Spanish, it was decided to use only pre-trained models in Spanish language. The pre-trained models selected, obtained from the *Huggingface* transformers library (https://huggingface.co/), were:

- *PlanTL-GOB-ES/roberta-base-bne* [10]. This model is based on the RoBERTa base model and has been pre-trained using the largest Spanish corpus known to date
- *davidmasip/racism* (https://huggingface.co/davidmasip/racism). This model is a RoBERTa model. It is used to predict whether a given text is racist or not
- *dccuchile/bert-base-spanish-wwm-uncased* [11]. This model (BETO) is a BERT Spanish version

For the first experiment, the models were trained with 10 epochs, 32 batch size, 128 token length, the optimization algorithm AdamW and a learning rate of 2e-5. Early stopping was used to avoid overfitting while training.

Table 3
Baseline results

Model	F1-score (class 1)	Accuracy	AUC ROC
RoBERTa	0.65	0.85	0.7580
racism	0.64	0.85	0.7531
BETO	0.57	0.84	0.7078

The results of this experiment can be seen in Table 3. We got these results as the baseline for the rest of our experimental framework with the aim of improving them using the techniques mentioned above. Furthermore, for the rest of the experiments, a searching for the best hyperparameters was carried out.

4.1. Data Augmentation

Data augmentation is a widely accepted technique for increasing the size of the training data by creating modified data from existing data. This approach is used when the initial dataset is too small to train, or when it is necessary to increase the number of examples of some classes to improve the performance of the model. Data augmentation can be carried out at different levels. At the word level there are methods such as synonym replacement, embedding replacement and language model replacement, among others. At the document level, the translation method can be highlighted.

For this task, the translation approach has been selected. This technique consists of converting the comment to a different language using an automatic translation model, and then converting it back to the target language. The use of this technique results in comments equivalent to the original but with different words. The models used were *opus-mt-es-de* (https://huggingface.co/Helsinki-NLP/opus-mt-es-de), which translates from Spanish to German, and *opus-mt-de-es* (https://huggingface.co/Helsinki-NLP/opus-mt-de-es) which translates the comment to Spanish again.

Data augmentation has been applied to the comments of the minority class, i.e., for each comment of class 1 an equivalent comment has been generated, as we can see in Table 4.

Table 4
Translation examples

Original comment	Comment generated
Este es uno de los problemas de inmigración incontrolada que ni siquiera tienes idea de quién está entrando.	Este es uno de los problemas incontrolados de inmigración que no tienes ni idea de quién va a entrar.
Creo también que quizá estaría justificado cerrar este centro y embalarlo en sus respectivos países.	También creo que podría estar justificado cerrar este centro y empaquetarlo en sus respectivos países.

The number of comments for each class after data augmentation can be seen in Table 5.

Table 5
Classes distribution

Dataset	Rows	Class 1	Class 0
Original (train/validation)	3053	697	2356
Augmented (train/validation)	3750	1394	2356

Data augmentation with substitution of synonyms using the *nlpaug* library (https://nlpaug.readthedocs.io/) and Easy Data Augmentation (EDA) [12] were also attempted, but good results were not obtained. Results achieved training the models with translation data augmentation technique are shown in Table 6.

Table 6
Data augmentation results

Model	F1-score (class 1)	Accuracy	AUC ROC
RoBERTa	0.66	0.84	0.7808
Racism	0.66	0.84	0.7808
BETO	0.62	0.82	0.7542

As can be seen in Table 6, the use of data augmentation techniques improved the F1-score of all three models. *RoBERTa* and *racism* models had an improvement of 1%, while the *BETO* model achieved an improvement of 5%.

4.2. Undersampling and Bagging

As mentioned in the previous sections, the dataset provided was significantly unbalanced. To address this problem, we also decided to adopt an UnderBagging approach. We first conducted some experiments to test whether the undersampling technique, applied to the majority class, improved the performance of the classifiers. To make the different samples, the number of positive examples was kept fixed, with the number of positive examples remaining the same as in the original dataset. Positive cases were sampled without replacement, so all of them were contained in each bootstrap. To determine the optimal number of examples of the majority class in the bootstrap samples, several tests with different sample sizes were performed. Finally, the best distribution between the examples of each class was 1:2.1, slightly more than twice as many examples for the majority class. Table 7 shows the distribution of the classes after undersampling.

Table 7
Classes distribution after undersampling

Dataset	Rows	Class 1	Class 0
Original (train/validation)	3053	697	2356
Undersampling (train/validation)	2167	697	1470

Five samples were extracted using random sampling with overlap for the majority class. The results achieved by the three models in each of the samples are shown in Table 8. The average value of the measurements is also shown. The results of *racism* and *BETO* models were slightly better when using undersampling than data augmentation. However, the *RoBERTa* model obtained worse results when undersampling was applied.

Table 8
Undersampling results

Model		F1-score (class 1)	Accuracy	AUC ROC
RoBERTa	S1	0.59	0.80	0.7401
	S2	0.22	0.65	0.4981
	S3	0.59	0.80	0.7386
	S4	0.59	0.79	0.7419
	S5	0.60	0.75	0.7729
	Average	**0.518**	**0.758**	**0.6983**
racism	S1	0.68	0.85	0.7931
	S2	0.68	0.85	0.7956
	S3	0.68	0.85	0.7983
	S4	0.66	0.82	0.7991
	S5	0.68	0.86	0.7869
	Average	**0.676**	**0.846**	**0.7946**
BETO	S1	0.65	0.80	0.7999
	S2	0.60	0.77	0.7642
	S3	0.61	0.78	0.7737
	S4	0.63	0.81	0.7769
	S5	0.63	0.81	0.7769
	Average	**0.624**	**0.794**	**0.7783**

To test the performance of the bagging meta-classifier, three and five bootstrap samples were combined. The three samples with the best results were chosen. Table 9 shows the results achieved using UnderBagging with three and five bootstrap samples.

Table 9
UnderBagging results

Model		F1-score (class 1)	Accuracy	AUC ROC
RoBERTa	Three samples	0.40	0.77	0.6195
	Five samples	0.44	0.79	0.6415
racism	Three samples	0.70	0.86	0.8027
	Five samples	0.68	0.85	0.7939
BETO	Three samples	0.63	0.79	0.7853
	Five samples	0.64	0.80	0.7982

4.3. Ensemble

Ensemble methods are some of the most advanced solutions to many machine learning challenges, typically in supervised machine learning tasks. These methods improve the predictive performance of a single model by combining the predictions of several models [13]. Each ensemble method requires a proper fusion of several learners to generate the final prediction model. The voting classifier estimator, created by combining different classification models, is a strong meta-classifier that balances the weaknesses of the individual classifiers [14].

In contrast to the UnderBagging described above, our hypothesis was that an ensemble of models trained on datasets using different balanced techniques could improve the final predictions. In this way, ensembles were constructed by combining models trained with datasets that were balanced using data augmentation and models trained with datasets that were balanced using undersampling. To build the ensembles, the models trained with data augmentation (*RoBERTa*, *BETO* and *racism*) described in Section 4.1, and the three samples meta-classifier UnderBagging using the *racism* model described in Section 4.2 were combined. The results obtained with the ensembles are shown in Table 10.

Table 10
Ensemble results

Ensemble model	F1-score (class 1)	Accuracy	AUC ROC
UnderBagging + RoBERTa + BETO	0.70	0.86	0.8152
UnderBagging + RoBERTa + racism	0.70	0.86	0.8213
UnderBagging + BETO + racism	0.69	0.85	0.8101

5. Results

In this section, we present the results of the different runs we have completed for subtask 1 of the competition. We used the official competition metrics (F-measure) to evaluate these results. We submitted three runs to the competition. The details of the modules and the differences between the three settings are described below.

- **I2C_III Run_1**. This submission comprises the predictions obtained by an UnderBagging meta-classifier using the pre-trained *racism* model. The meta-classifier consisted of three classifiers. Each classifier was trained on a sample from our training dataset to which undersampling was applied with a ratio of 1:2.1. Majority voting was used to obtain the final prediction on the test file. To select the three best models, they were tested with our labelled test dataset.

- **I2C_III Run_2**. This submission is similar to the first one. In this case, the UnderBagging meta-classifier consists of five *racism* models trained with random samples from the full training dataset provided by the organizers.
- **I2C_III Run_3**. Predictions for the third submission were obtained using a meta-classifier ensemble of three classifiers. One of them was the UnderBagging classifier used in the first run, and the other two were the *RoBERTa* and *BETO* models trained with the full training dataset provided balanced by data augmentation.

A search for identifying the best hyperparameters was carried out during the experimentation phase. We implemented the following hyperparameters for the three runs: learning rate as 5e-05, batch size as 16, 128 as token length, weight decay as 0.01, the optimization algorithm AdamW, and maximum epoch as 5.

Our results in the competition for subtask 1 among the participants (Table 11) show the success of our proposed model achieving the first place in the ranking.

Table 11
Ranking of participants' systems in subtask 1 of DETESTS shared task

Ranking	Run	F-measure
1	I2C_III Run_1	0.7042
2	I2C_III Run_2	0.7038
3	I2C_III Run_3	0.7032
10	Lak_NLP Run_4	0.6627
83	TMNT Run_1	0.4866
-	FastText+SVC	0.4860
146	TMNT Run_2	0.0241

6. Conclusions

This paper presents the participation of the I2C research group at the DETEction and classification of racial STereotypes in Spanish shared subtask 1 at IberLEF 2022. Our proposal demonstrates that the use of pre-trained linguistic models, meta-classifiers and data balancing techniques can help to obtain excellent results in a binary text classification task. The three proposed architectures obtained the first three positions in the ranking, demonstrating their validity in detecting stereotypes in comments related to immigration.

The use of UnderBagging was shown to be a very effective approach to solving the task. The ensemble of trained models with different data balancing approaches also showed a successful performance.

7. References

[1] A. Ariza, W.S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of the DETESTS Task at IberLEF-2022: DETEction and classification of racial STereotypes in Spanish, Procesamiento del Lenguaje Natural, 69, 2022.

[2] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, W. Daelemans, The automated detection of racist discourse in dutch social media. Computational linguistics in the Netherlands journal, 2016, 6, 3-20.

[3] F. Rodríguez, J. Carrillo de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access, 2020, 8, 219563-219576.

[4] L. Kaati, A. Shrestha, K. Cohen, S. Lindquist, Automatic detection of xenophobic narratives: A case study on Swedish alternative media. In 2016 IEEE conference on intelligence and security informatics (ISI), pp. 121-126.

[5] P. Mishra, M. del Tredici, H. Yannakoudakis, E. Shutova, Abusive language detection with graph convolutional networks, 2019, arXiv preprint arXiv:1904.04073.

[6] Y. Gaci, B. Benatallah, F. Casati, K. Benabdeslem, Masked Language Models as Stereotype Detectors? In EDBT 2022.

[7] J. Sánchez-Junquera, P. Rosso, M. Montes, B. Chulvi, Masking and BERT-based Models for Stereotype Identication, 2021, Procesamiento del Lenguaje Natural, 67, 83-94.

[8] M. Bayer, M.A. Kaufhold, C. Reuter, A survey on data augmentation for text classification, 2021, arXiv preprint arXiv:2107.03158.

[9] B.S. Raghuwanshi, S. Shukla, Class imbalance learning using UnderBagging based kernelized extreme learning machine, 2019, Neurocomputing, 329, 172-187.

[10] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C.P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A.Gonzalez-Agirre, M. Villegas, MarIA: Spanish Language Models, 2022, Procesamiento del Lenguaje Natural, 68.

[11] J. Cañete, G. Chaperon, R. Fuentes, J.H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, PML4DC at ICLR 2020.

[12] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019, arXiv preprint arXiv:1901.11196.

[13] O. Sagi, L. Rokach, Ensemble learning: A survey, 2018, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

[14] I.E. Livieris, L. Iliadis, P. Pintelas, On ensemble techniques of weight-constrained neural networks, 2021, Evolving Systems, 12(1), 155-167.