Estimating Ground Truth in a Low-labelled Data Regime: A Study of Racism Detection in Spanish

Paula Reyero Lobo, Martino Mensio†, Angel Pavon Perez†, Vaclav Bayer†, Joseph Kwarteng†, Miriam Fernandez, Enrico Daga, Harith Alani

† Authors contributed equally

Knowledge Media Institute, The Open University, UK

{paula.reyero-lobo, martino.mensio, angel.pavon-perez, vaclav.bayer, joseph.kwarteng, miriam.fernandez, enrico.daga, harith.alani}@open.ac.uk

Abstract

Obtaining reliable and quality training datasets is resource-intensive, especially in interpretation and human judgment tasks, such as racism detection. Related work reveals that annotators subjected to hate are more sensitive to labelling something as offensive and advocate giving more voice to these collectives. This study analyses a new dataset for detecting racism in Spanish, focusing on solving a ground truth estimate given a few labels and high disagreement. Most annotators may not have previous experience with racism, as only three belong to the Black community. Our empirical results show better performance at lower thresholds for classifying messages as racist, which may be due to how annotators being permissive in identifying racist content propagates to the model. This analysis can be crucial for tailoring a general model to the specific needs of a particular individual or group. Especially in applications such as online abuse, detection models that reflect the viewpoint of crowdworkers may not be sufficient to detect all the intricacies of these social challenges.

Introduction

Automatic detection of racism, hate speech, and similar online abuse is challenging, given its subjective nature, as the perceptions of people exposed to such content can vary based on their demographics and previous experiences (Garg et al. 2022; Sang and Stanton 2022). However, the vast majority of resources available to combat online harm are constructed using crowdworkers or "expert judges" who have been trained for the task and lack information about the annotators' relationship to online abuse (Poletto et al. 2021). In addition, knowledge about racism in non-English is scarce, so data collection in other languages is essential as they express these phenomena in specific ways. In these cases, obtaining quality labels can be more challenging, as they tend to have fewer resources allocated to them (Field et al. 2021).

This paper focuses on solving the estimation of ground truth annotations for detecting racism in Spanish under these limitations. Specifically, the dataset contains a low number of labels from a small group of volunteer annotators and a notably high disagreement scenario. We find many messages only seen by one annotator or given different labels by the same person. An example of the latter is the message *"8-year-old girl killed in Pakistan for letting parrots escape from the family she worked for as a domestic servant - Wellcome to Europe!"*. The only annotators who saw the message twice assigned it contradictory labels.

Related work investigates the impact of annotator demographics (Gordon et al. 2022; Kumar et al. 2021) or crowdsourcing design (Vidgen et al. 2021) for ground truth estimation in complex annotation tasks (i.e., of texts that need human interpretation and judgement). To our knowledge, this is the first work to investigate the impact of different ground truth estimation methods on the ability to detect racist messages despite limited resources and a lack of annotator demographics. Our empirical results show that the models perform better when using a lower threshold to classify a message as racist, which could be due to how the permissiveness in annotating racist messages may propagate to the model. We release the code for this work at: *https://zenodo.org/badge/latestdoi/487601308*

Related Work

Some factors significantly influence the labelling of abusive texts, which may prevent consistency between annotators with different backgrounds (Sang and Stanton 2022). Groups historically at risk of abuse are more likely to notice these messages (Kumar et al. 2021; Olteanu, Talamadupula, and Varshney 2017). As a result, toxicity scores from popular detection models may only reflect views of specific identities or groups. For example, scores aligning more with White's perspective on African American English (AAE) toxicity than Black annotators (Sap et al. 2021).

Recent work aims to add this dimension to the learning model. Using only a few of the demographic attributes as a feature in model training improves significantly the quality of recognition (Kocoń et al. 2021). Ensuring equitable representation of specific demographic groups addresses the instability of standard labelling approaches (Gordon et al. 2022). However, most resources do not contain demographic information (Poletto et al. 2021). Several rounds for correcting annotation errors and showing the model the most controversial cases help deal with the complexity of subjectivity in a crowdsourcing task (Vidgen et al. 2021). This paper considers the case of ground truth estimation under low annotation and consequent high disagreement. These

challenges add to the complexity of racist discourse annotation (Aroyo et al. 2019), which may occur in resources for under-researched linguistic contexts (e.g., in Spanish datasets (Basile et al. 2019)).

Data and Methods

We conduct an empirical study to estimate the ground truth of racism in a low-labelled data regime. First, we present an exploratory analysis to understand the annotation problem better. This analysis leads to the posterior experimental design for ground truth estimation. Finally, we analyse the impact of different estimation methods on detection.

Data Description

This study used a dataset provided by BCNAnalytics for the Datathon Against Racism (BCNAnalytics 2022). Data were collected from August 2020 to January 2022 from Twitter. It contains 9291 labels for 5672 unique messages in the training set and 59 samples for evaluation. The dataset documentation provides the annotation guidelines and a list of the 24 *volunteer* annotators' names: nine were members or had links with the organising committee, ten were employed by sponsoring companies, and five became interested through social media. Each annotator has a unique identifier and could annotate the same message more than once. The annotation followed a mutually exclusive labelling scheme with 4641 "non-racist", 4229 "racist", and 421 "unknown" labels.

Annotation Assessment

We characterise the problem of ground truth estimation of the dataset at the message and label levels. First, we focus on how the content was assessed. If we look at the distribution of labelling, messages were seen by few annotators (see Table 1). Most messages have one label, and only a small fraction have more than five labels.

No. Labels	Racist	Non-racist
1	59 %	59 %
2	28 %	30 %
3	8 %	5 %
4	3 %	2 %
5	1 %	3 %
> 5	1 %	1 %

Table 1: Percentage of messages in the Racist and Non-racist samples with increasing number of labels (No. Labels).

Furthermore, we investigated the agreement reached by the annotators. We consider the mean labels by the number of annotators and their inverse to compare the percentage of agreement on racist and non-racist messages, respectively (Figure 1). For example, in cases where only one annotator gave two different labels (e.g. 1 and 0), the agreement score for the message would be 0.5. The figure shows the average of these scores in increasing annotator number. As expected, agreement decreases with increasing annotators. Interestingly, this tendency fluctuates more in the non-racist sample, indicating that annotators had difficulties, especially when

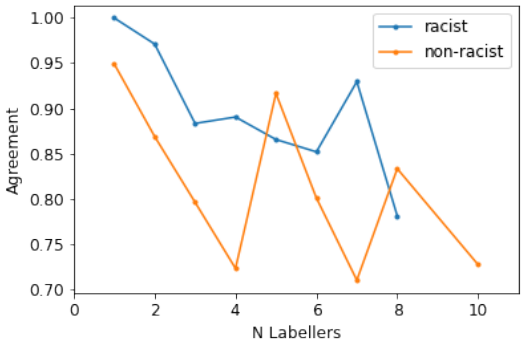

Figure 1: Agreement in racist and non-racist messages at increasing annotator number (N Labellers).

dealing with these texts. Cohen's kappa coefficient(κ) supports our findings, as they are in the range of poor to slight agreement (Landis and Koch 1977). Particularly, $\kappa = 0.010$ ($\sigma = 11.46$) for "non-racist" labels, $\kappa = 0.011$ ($\sigma = 13.02$) for "racist", and $\kappa = 0.002$ ($\sigma = 1.75$) for "unknown".

Ground Truth Estimation

In this context, it is essential to model individual annotations so that the model learns significant racist features. First, we have to find suitable coding for the labels. Thus, "unknown" lies in the middle between the two extremes of the racism scale (i.e. where completely racist is 1.0), so we assign it 0.5. Second, we consider a baseline and two methods for aggregating the individual labels in each message:

- The baseline (raw) includes all the individual annotations excluding the "unknown" labels.
- The majority vote (`m_vote`) takes the average of the labels for each text.
- The weighted majority vote (`w_m_vote`) adds a weight reflecting the number of times the message was labelled.

We assign the weight by doing a data exploration of sampled messages in different ranges (i.e. very racist, intermediate, and not racist). We assign a value sufficient to give more certainty to the majority vote the more labels without compromising the value of the vote (e.g., giving too high priority to the few messages annotated more than five times). For this reason, we cannot assign a too high value to the weighting of the label number, given the limited number of possible times messages were annotated. Thus, we compare different values and investigate using a weight of 15%, where the weighted vote for a text with n tags would be:

$$\texttt{w\_m\_vote} = 0.85 * \texttt{m\_vote} + 0.15 * \texttt{n\_annot} \quad (1)$$

where `n_annot`=n/max if `m_vote` is "racist", and `n_annot`=$(1 - n/max)$ otherwise, and max is the maximum number of labels. Giving a higher probability to messages seen by more annotators ("the more eyes, the better") may be a determining factor given the low-annotation conditions in our corpus.

Model	Test Set	Fine-tuning Epochs							
		1		2		3		4	
Raw	Validation	0.790		0.819		0.790		0.799	
	Evaluation	0.867		0.842		0.905		0.888	
	Average	<u>0.828</u>		0.830		**0.848**		0.844	
		S	NS	S	NS	S	NS	S	NS
Majority vote (m_vote)	Validation	0.735	0.799	0.771	0.825	0.734	0.830	0.767	0.841
	Evaluation	0.884	0.857	0.867	0.827	0.857	0.825	0.851	0.866
	Average	0.810	0.828	0.819	0.826	<u>0.795</u>	0.828	0.809	**0.854**
Weighted majority vote (w_m_vote)	Validation	0.780	0.846	0.834	0.836	0.845	0.841	0.821	0.828
	Evaluation	0.867	0.813	0.852	0.900	0.847	0.881	0.888	0.896
	Average	<u>0.824</u>	0.830	0.843	**0.868**	0.846	0.861	0.855	0.862

Table 2: F1 scores of the models trained with different ground-truth estimates. We compare the *strict* (S) and *non-strict* (NS) version of the models with aggregated labels. We report the set scores for the validation (568 aggregated and 862 raw samples), the evaluation (59 samples) and their mean. We show in bold the best candidate models and underline the lowest scores.

Impact Evaluation on Racism Detection

We conduct a comparative analysis of the performance, error and bias of models trained with different ground truths, i.e., using individual (raw) or aggregated annotations. We use BETO-uncased (Cañete et al. 2020), a pre-trained linguistic model for Spanish with the HuggingFace transformer library (Wolf et al. 2019). We perform four epochs of fine-tuning, as suggested in Devlin et al. (2018), and compute performance in each epoch.

Due to the small number of samples in the evaluation sample (i.e. 59), we randomly draw 10% from the training set and use it as a validation set. Accordingly, we compare the performances of the models on a larger validation sample following their corresponding ground truth estimation method (i.e. with 568 aggregated and 862 raw samples) and on the smaller evaluation set given in the competition.

We evaluate the models considering their method of label aggregation and treatment of borderline examples for the binary classification of racism messages. In the case of aggregated labels, we consider that the inclusion of borderline messages (i.e., 0.5) as non-racist results in a *strict* classifier and vice versa (i.e., *non-strict*). That is, we take into account different thresholds of the model to classify a message as racist: the higher the racism detection threshold, the stricter the model.

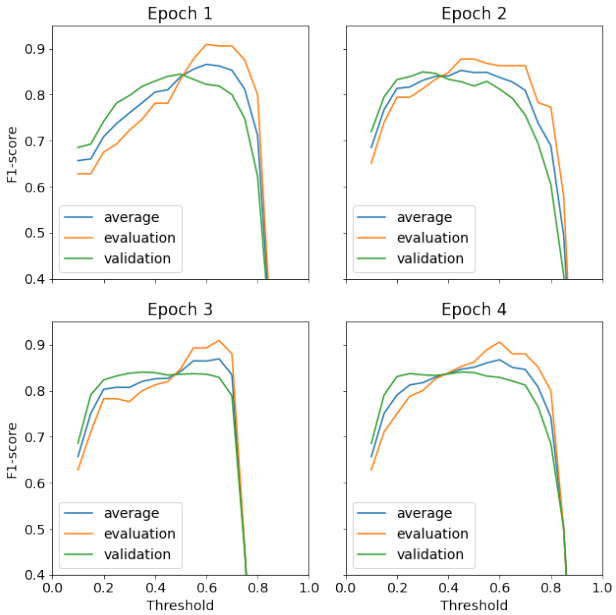

Figure 2: F1-score of the *non-strict* model trained on the weighted majority vote labels for racism detection at different thresholds.

Results

We present our results in the line of two main findings:

- Having few annotators from the racism target group may lead to a more permissive model of racist messages.
- The ground truth estimation plays an influential role in reducing annotator bias.

The *non-strict* version of the model trained with the weighted majority vote obtains the highest scores (Table 2), outperforming all models in all epochs. In particular, its F1 score is $\sim$ 3% higher relative to the baseline (from 0.848 to 0.868). This increase is relevant given the drop in aggregated labels (i.e., m_vote). The raw and m_vote models share a pattern of high performance mainly in the evaluation sample (i.e. of 59 records). We observe that both versions of weighted voting obtain more balanced scores in the larger validation sample, which is a good indicator of robustness and favours the value of the number of labels for ground truth estimation. More importantly, we find that the *non-strict* versions of the models outperform their counterparts, which supports our first finding. In fact, the *strict* model scores are worse than the baseline in the case of the majority vote and only comparable in the weighted model. We argue that the low threshold for classifying a message as racist may be due to annotators being permissive in identifying racist content, leading to underestimating racism.

For models with the lowest threshold, messages will only be "non-racist" when the maximum number of annotators

Model	FP	High Tox	FN	Low Tox
raw	3.39%	0%	5.09%	**1.7**%
m_vote	11.86%	0%	1.70%	0%
w_m_vote	10.17%	**1.7**%	0%	0%

Table 3: Percentage of false positive (FP) and false negative (FN) errors in the best candidate models. The right-side column of each sub-table shows the percentage of errors with toxicity higher than 80% or lower than 20%, respectively.

agree on it. We observe the performance of the w_m_vote at different thresholds and find that models obtain the best results close to the lowest threshold in all epochs (Figure 2). Stricter models (i.e. with a higher threshold for defining a message as "racist") may be overfitting on the smallest sample given for evaluation.

Given the limitations of performance metrics for analysing the quality of machine learning models, we ask how much better or worse each model's best candidate is. Our goal is to understand better where their differences come from by comparing their errors (Table 3). First, the model is more prone to both miss potentially racist content and identify it incorrectly when trained on individual annotations. The leap in false positives when resorting to the crowd's wisdom to determine the ground truth is striking and supports the conclusion of the exploratory analysis that the "non-racist" sample was more challenging to annotate. However, the weighted voting reduces false positives and appears to have better recall of racist messages so that it may be more "protective" of vulnerable groups.

However, as this analysis reflects the annotations we intend to investigate in the first place, we rely on an external resource to check the ground truth. Using the Spanish version of the Perspective API, we find false positives with high toxicity only in the weighted model (Google and Jigsaw 2017). Specifically, this is a 93% toxic message but annotated as "non-racist": *"Another buffoonery from the Colombian thief and drug criminal @NicolasMaduro Moros."* Similarly, the one case found of false negatives with low toxicity (i.e. 14%) in the raw model is another example of implicit bias: *"Arrested 12 Moroccan stowaways who reached Motril (Granada) on a ferry from Melilla"*. Racial bias in newspaper reporting is a well-known social problem (Teo 2000; Sonnett, Johnson, and Dolan 2015), which occurs due to the consistent mention of specific demographic groups in criminal news. Our findings show that these examples can be complex for labellers to annotate and propagate even in larger models such as Google's Perspective API.

Bringing model diagnostics a step further, we uncover not observable assets in the performance score. Further evaluation helps to elucidate the error reasons, which in some cases are hampered by a ground truth that may not be as expected. Our results confirm the impact of estimation for learning meaningful features of racism, even if these sometimes do not align with the ground truth due to the annotator bias that we aim to overcome.

Limitations

The dataset used for this analysis was published as part of a competition and has not been peer-reviewed. Our analysis focused on the specific problem of the impact of ground truth estimation on the detection of racism and allowed us to uncover informative evidence of annotator bias. However, we note that further exploration of data collection issues would be necessary to validate the utility and scope of this dataset in detecting racism messages.

Second, we had limited information on annotator demographics as we only knew their race. Our results are consistent with recent findings on the impact of specific identities and beliefs on the underestimation of toxicity (Sap et al. 2021). Having only three Black annotators could be why the detection threshold for racist messages is low. However, we need a larger sample size and annotator traits to validate the nature of this problem.

We support the inclusion of annotators from targeted groups, as they can capture the nuances of abuse (Curry, Abercrombie, and Rieser 2021). Inclusion in data collection and tagging may be the only way to capture the content that marginalised communities find most harmful (Maronikolakis et al. 2022).

Conclusion and Future Work

We focus on detecting racism in Spanish messages to show the importance of the label origins in constructing a reliable training dataset. Using the state-of-the-art model for racism detection in Spanish (Benítez-Andrades et al. 2022), we show that our models perform best while selecting the low threshold for classifying a message as racist. Moreover, estimating ground truth from these sparse annotations is crucial when human interpretation and judgment are a source of variability, and exploiting them in the "worst-case" scenario may open doors to large-scale cases (e.g. crowdsourcing).

This analysis may be critical to ensure a *personalised tuning* tailored to specific demographic groups' needs (Kumar et al. 2021). These models appear to be strong candidates for future work due to the biases detection models exhibit against vulnerable demographic groups (Hutchinson et al. 2020), and AAE in particular (Kim et al. 2020; Sap et al. 2019).

Since these complex messages are difficult to be learned with detailed instructions or label definitions, it is essential to focus the work around the people in the communities who suffer from this form of hate and bias of the NLP systems and involve them in the annotation process (Blodgett et al. 2020).

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS-Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, 1100–1105.

Basile, V.; Bosco, C.; Fersini, E.; Debora, N.; Patti, V.; Pardo, F. M. R.; Rosso, P.; Sanguinetti, M.; et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, 54–63. Association for Computational Linguistics.

BCNAnalytics. 2022. Datathon Against Racism. https://bcnanalytics.com/datathon/.

Benítez-Andrades, J. A.; González-Jiménez, Á.; López-Brea, Á.; Aveleira-Mata, J.; Alija-Pérez, J.-M.; and García-Ordás, M. T. 2022. Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. *PeerJ Computer Science*, 8: e906.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

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

Curry, A. C.; Abercrombie, G.; and Rieser, V. 2021. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. *arXiv preprint arXiv:2109.09483*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 1905–1925. Online: Association for Computational Linguistics.

Garg, T.; Masud, S.; Suresh, T.; and Chakraborty, T. 2022. Handling Bias in Toxic Speech Detection: A Survey. *arXiv preprint arXiv:2202.00126*.

Google; and Jigsaw. 2017. Perspective API. https://www.perspectiveapi.com/.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J. T.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *arXiv preprint arXiv:2202.02950*.

Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501.

Kim, J. Y.; Ortiz, C.; Nam, S.; Santiago, S.; and Datta, V. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv e-prints*, arXiv–2005.

Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; and Kazienko, P. 2021. Offensive, Aggressive, and Hate Speech Analysis: From Data-centric to Human-centered Approach. *Information Processing & Management*, 58(5): 102643.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security*, 299–318.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Maronikolakis, A.; Wisiorek, A.; Nann, L.; Jabbar, H.; Udupa, S.; and Schuetze, H. 2022. Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments. *arXiv preprint arXiv:2203.11764*.

Olteanu, A.; Talamadupula, K.; and Varshney, K. R. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the 2017 ACM on Web Science Conference*, 405–406.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2021. Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. *Language Resources and Evaluation*, 55(2): 477–523.

Sang, Y.; and Stanton, J. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *International Conference on Information*, 425–444. Springer.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Sonnett, J.; Johnson, K.; and Dolan, M. 2015. Priming Implicit Racism in Television News: Visual and Verbal Limitations on Diversity. *Sociological Forum*, 30.

Teo, P. 2000. Racism in the news: A critical discourse analysis of news reporting in two Australian newspapers. *Discourse & society*, 11(1): 7–49.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1667–1682.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.