DaMinCi at IberLEF-2022 DETESTS task: Detection and Classification of Racial Stereotypes in Spanish

David Cabestany[1,2,†], Clara Adsuar[2] and Miguel López[2,†]

[1] *University of Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain*

[2] *University of the Basque Country, Manuel Lardizabal pasealekua 1, 20018 Donostia, Spain*

Abstract

In this paper we present our participation in the first Subtask of the DETESTS task -DETEction and classification of racial STereotypes in Spanish-, which took place as part of IberLEF 2022, the 4th Workshop on Iberian Languages Evaluation Forum at the SEPLN 2022 Conference. We first describe the corpus provided by the organizers, which consists of two domain-specific subcorpora: NewsCom-TOX and StereoCom. We then present the four approaches we developed to solve the addressed task, of which three are BERT-based (2 RoBERTa models fine-tuned in similar tasks and 1 RoBERTa with Adapters) and another one is a fine-tuned GPT2. Regarding the results, we obtained an F-score score of 0.659.

Keywords

IberLEF2022, DETESTS, multi-label classification, racial stereotypes

1. Introduction

In this work, we have participated in the IberLEF 2022 competition, more specifically on the task known as "DETEction and classification of racial STereotypes in Spanish (DETESTS)" [1]. The purpose of DETESTS is to develop a system that is capable of detecting the presence of stereotypes -targeting social or ethnic groups- on user-generated online content. The task itself consists of two subtasks: a binary classification one as well as a multilabel classification problem. We have focused exclusively on the first Subtask of the two proposed by the organizers. The data used in this project is the one provided to the contestants, which is conformed by two different corpora: NewsCom-TOX and StereoCom.

Along this project, we will explain the ways we confronted the data that was given for the task, as well as how we dealt with the issues that arose and our decision of implementing four pipelines to solve the chosen subtask. Being all of them Transformer-based models, we decided

IberLEF 2022, September 2022, A Coruña, Spain.

[†] These authors contributed equally.

d.cabestany@ub.edu (D. Cabestany); claraadsuar@gmail.com (C. Adsuar); miguelopezotal@gmail.com (M. López)

https://github.com/dcabesma/ (D. Cabestany); https://github.com/cladsu (C. Adsuar); https://github.com/miguelopezotal (M. López)

0000-0002-0122-7931 (D. Cabestany); 0000-0003-4093-2307 (C. Adsuar); 0000-0002-4854-8062 (M. López)

to implement two fine-tuned RoBERTa, one RoBERTa with adapters, and a GPT2 fine-tuned for multilabel sequence classification.

2. Task Description

The aim of the DETESTS task, as proposed by [1], is to detect and classify stereotypes in sentences from comments posted in Spanish. These comments were selected among the responses to online news articles regarding immigration.

The proposed task we have chosen to develop is the Subtask 1. The core of this task is a binary classification of stereotypes; that is, its main objective is to develop a classification model able to discern whether there is at least one or more stereotypes present in an online comment. Therefore, if a sentence from an instance of the given dataset for the task showed at least one stereotype, from among the 13 possible types of stereotypes the dataset provided as features, a model would classify the sentence as positive (marked as "1", that is, having a stereotype or several). On the other hand, if a sentence didn't present any stereotypes it would be classified as negative (marked as "0").

3. Dataset

The dataset used for this project is formed by two corpora: one of them is called NewsCom-TOX [2] and the other one receives the name of StereoCom. Both corpora incorporate online comments that originated as user responses to online news articles, extracted from two main sources: Spanish newspapers (e.g. ABC, elDiario.es, El Mundo, etc.) and discussion forums (e.g. Menéame). The comments leveraged for these two corpora correspond to a specific timeframe: in the case of NewsCom-TOX, the set of articles dates from August 2017 to August 2020, while the articles from StereoCom corpus are dated from June 2020 to November 2021. Regarding the number of sentences from each corpus, the dataset encloses a total of 5,269 sentences:

- 3,306 sentences from NewsCom-TOX
- 2,323 sentences from StereoCom

On average, the 24% of the total amount of data shows at least one stereotype.

The corpora authors manually selected the articles based on some criteria, mainly the controversial subject matter, the potential toxicity of the messages as well as the number of published comments (at minimum 50 comments). They also did a key-word search for finding articles to build the corpora, settling on those that matched the keywords of racism and xenophobia. These articles previously mentioned conform the highest percentage of instances in the corpora where stereotypes can be found.

The comments collected for building the corpus are divided into several sentences, which are individually placed on their own distinct dataset instances. Each dataset instance includes, alongside its sentence, some additional metadata that relates that sentence to both the comment it was taken from and to the position of that sentence in the original comment. Additionally, each dataset instance contains a one-hot vector of features that tell whether the specific instance has one type of stereotype -for instance, whether a stereotype present is targeted at a specific ethnicity ("racial_target"), if there is a stereotype that is implicit in the message ("implicit") or if it contains dehumanisation. There is a total of 14 features, and each one has a value of either 0 (False) or 1 (True), which denotes whether that specific kind of stereotype is present or not. Among these features, we were specially interested in that under the label name of "Stereotype", which tells whether the specific dataset instance contains any stereotype at all. If any of the other features has a value of 1 (True), this feature will also have a value of 1 -meanwhile, it will have a value of 0 if the rest of the features happen to be 0. For the purposes of our proposal, in which we have participated exclusively on the first Subtask of the DETESTS competition, we have mostly relied on the "Stereotype" label, but one of our models has also used the remaining labels for its classification task.

The division of the dataset provided by the organization team is 70% for the training set and the other 30% is for testing, and thus it is hold by the organizers to test our models. From the provided training set to the contestants we saved 10% for development.

3.1. Annotation Scheme

The growing population in social media has led to the development of many tasks in order to address and diminish the hateful speech on the Internet. Stereotypes are hard to identify and classify due to the fact that many times are expressed implicitly. Therefore, we first set out to understand the concept of stereotype and its linguistic features.

As it is defined in the DETESTS competition's webpage: "A stereotype is defined in social psychology as a set of beliefs about others who are perceived as belonging to a different social category".

Due to our historical and cultural background, some stereotypes could be considered "unoffensive" generally due to the usage they have had along History. We can assume that many times people do not express stereotypes explicitly (emphasized by insults or swearwords that could lead to a clear classification of the stereotype itself), but in an implicit manner (through irony and sarcasm).

We should take into consideration that common words that help us target the topic are not inherently negative. Therefore, we should focus on the criterion of homogenization in the comments we are studying. In DETESTS' webpage, we can find the definition of the term homogenization, that is described as "a process of the generalization of a feature to the status of a social category, which negates individual diversity".

In this work, the attributes for the stereotype classification that we address are divided into 13 labels, those are the *racial_target, other_target, implicit, stereotype, xenophobia, suffering, economic, migration, culture, benefits, health, security, dehumanisation,* and the label *others* for the unspecified stereotypes.

4. Models

For tackling Subtask 1 of the DETESTS task of the IberLEF 2022 competition, we decided to follow an approach where we developed four classifiers with the Transformer architecture [3]. More precisely, one of them is based on the GPT-2 architecture [4] and the remaining ones on

RoBERTa [5]. We therefore made use of a neural-based approach to the binary classification problem as posed by the Subtask 1 of the DETESTS task.

We handled the construction of these classification models following different approaches, but they all had in common the fact that they benefited from models already pre-trained on Spanish text data. The use of Spanish pre-trained models was necessary since the Task relied on a Spanish dataset.

Our models followed these different building approaches:

- Fine-tuning approach on a large pre-trained model on Spanish data.
- Fine-tuning an already fine-tuned model on a closely related domain -as an experiment to see its efficiency and possible shortcomings.
- Other different training techniques such as Adapters [6].

In the following sections, we will explain in detail the development of the models.

4.1. GPT2 for Multilabel Sequence Classification

For this work, we used the pre-trained OpenAI GPT2-medium [4] allocated in OpenAI for Sequence Classification with multi_label_classification problem-type configuration, this model consists of 1024 neurons per hidden layer, 16-heads, and 345M parameters. In general, GPT models are well-known models widely used for text generation tasks, but we used the medium GPT2 as an attempt to perform binary classification prediction.

The Sequence Classification function uses the last token to do the classification, as other causal models do, such as GPT-1. Thus, GPT models require knowing the position of the last token for learning. The way of knowing this last token is defining a padding token (pad token) in the configuration of the transformer. If we define a pad_token_id in the configuration, the model finds the last token that is not the pad_token in each row. If we do not define the pad_token, it only considers the embed value of the last token in a batch row. Since it cannot guess the padding tokens when inputs_embeds are passed instead of input_ids, the last token of the input sequence contains all the information needed in the prediction. Taking this assumption into account, we can use that information to predict a classification task instead of a generation task [7].

In other words, instead of using the first token embedding to predict as BERT (right padding), we use the last token embedding to predict with GPT2 (left padding).

We configured the Sequence Classification model parameters for reproducibility as follows, problem_type configured as a "multi_label_classification" we had to change the output features of the GPT-2 to 13 with the parameter num_labels = 13. The seed is 32; we used ten epochs for training with a batch size of 1 and a max length of 60. We also configured a Decision Threshold to 0,60 because it was the threshold with the best performance in the development set.

The batch size, in this case, is equal to one due to Colab memory limitations. Once we set the optimum batch size for this model, we tested it to see the maximum length it supports to ensure the model works with the most number of words simultaneously for every dataset comment.

The optimizer is Adam algorithm with weight decay fix [8] for grouped parameters with a correction of the bias, and a Learning Rate of 2^{-5}.

As we wanted to ensure that the model learns about the labels with less support, we made a simple oversampling extending all the labels that only occurs once in the original dataset to confirm that all the sentences are included in the development and the training set. After that, we randomized the dataset to split it into train and development sets. Moreover, we deleted the previous oversampled occurrences of the development set.

4.2. RoBERTa-Based Models

4.2.1. Fine-Tuning Already Fine-Tuned Models

In this section, we will describe the RoBERTa-based models we used in order to deal with the task. These models were already fine-tuned for specific domains. The domains are Racism and Online Harassment on Internet comments and posts. We consider the domains previously mentioned as closely related to the topic of the competition. For instance, one of the already fine-tuned models we examined was specialized in detecting racism on online user-generated comments, something that is not very far away from the objective of detecting the presence of different sorts of stereotypes on a text of a similar nature.

Our approach had the aim of accomplishing Transfer Learning. For that reason, we reused the weights of a fine-tuned model as the starting point for our task. As a matter of fact, other authors [9] have already proposed similar frameworks based on performing fine-tuning on already fine-tuned models.

Using this approach, we trained two classifiers, both having an underlying RoBERTa architecture pre-trained on Spanish data.

The first model [10] consists on a fine-tuned model that was taken from an original model trained for detecting if a given text is racist or not, provided via HuggingFace. The original model was fine-tuned on a RoBERTa-BASE model pre-trained on data from Biblioteca Nacional de España. This model was part of a competition called "Datathon against Racism" [11], held by BCN Analytics. The model's authors have not provided any written reference to their work yet as of the writing of this paper, which can be attributed to the competition being held in a close time frame to the making of this article.

The second model [12], on the other hand, was based on a model that had been fine-tuned on a domain not as closely related to ours, but still similar. The original model detected online harassment on Internet comments and posts, a task that we have assumed can be relatively well adapted to the task of being fine-tuned for the detection of stereotypes of different sorts. The original model was leveraged from HuggingFace, and its authors claim their original model achieved a 0.9167 accuracy score on their test set. The original model was also part of another competition [13], held by the organization 'SomosNLP'. We can also notice the lack of another written article for this model, which could be attributed to the relatively recent organization of this event.

Both original fine-tuned models had a classification layer whose main output labels were either "0" or "1", which means they were designed for a binary classification task, which is perfect for our purpose. We decided to keep the original classification layers of both models and retrain their weights using the provided dataset for the DETESTS SubTask we participated in. We opted to keep the classification layers from the original models -instead of discarding them- since we assumed that the original models, having been fine-tuned on similar domains to ours, would have a benefit on performance accordingly. Although the original fine-tuned models are binary classifiers, their output was not exactly either "0" or "1" [10] exported labels ’LABEL_0’ and ’LABEL_1’ and [12] did the same with ’No acoso’ (Spanish for ’No harassment’) and ’Acoso’ (’Harassment’). Despite this slight difference in the output, we could map it into our own labels system.

We chose not to manipulate any features from our dataset, other than the "0" or "1" value given under the "stereotype" feature in each training instance. This decision was made in order to avoid incompatibility issues with the already fine-tuned models. This meant not using in this case the variety of features provided in the training set, such as if there is a ’racial target’ in a specific instance or if there is an example of ’dehumanisation’ of a specific group or culture.

In both cases, we did a split on our training data, and we assigned 10% of it for development during fine-tuning. We also used the same hyper-parameters for fine-tuning both models, this is a seed of 123, we used 10 epochs with a batch size of 32. In this case we do not need to limit the length of the text. The device we used is a GPU in cuda. The optimizer is the Adam algorithm with weight decay fix with a learning rate of 5^{-6}.

4.2.2. Adapter-Based Model

Adapters [6] is a training technique, with a similar objective to fine-tuning, but which has a lower computational overhead and obtains comparable results to traditional fine-tuning. It consists on a series of bottleneck layers -with a new set of trainable weights- that are inserted between the layers of a pre-trained model. During training, only the weights of these layers are changed, while the original model’s weights remain constant.

We decided to adopt this training approach in order to train a RoBERTa BASE model that had been pre-trained on Spanish data stemming from Biblioteca Nacional de España [14], whose corpus is the largest Spanish source of texts obtained by web crawling.

The reason behind the use of this architecture was to experiment on a base, pre-trained RoBERTa model with the given dataset, avoiding the overhead of extensive GPU runtime, as well as the possibility of obtaining increasingly better results compared to mere fine-tuning.

We made use of a special library dedicated to the training of Adapters on Transformer-based architectures, called AdapterHub [15]. This framework allowed us to easily insert a series of Adapter layers inside the model, as well as adding a binary classification head on top of the original model. The framework also took care of only training the weights of the Adapter layers; as well as those of the added classification layer. For training the model, we opted for a Learning Rate of 1^{-4}, batch size of 32 and a total of 6 epochs. We did a split on the training set, where we devoted 20% of it for development.

5. Results

As the organizers of the task describes on the webpage, the result’s evaluation of the competition are made following the task 12 from SemEval-2021 [16], so the results are evaluated using F1 classification metric and the cross-entropy between the system output values and the soft labels generated by a probabilistic normalization procedure.

In this section we want to compare the F1-score from the development set with the F1-score of the test set provided by the organization team, as reflected in Table 1.

Table 1
Comparison of the obtained F1-score results on the development set with the test set provided by the organizers.

Model	Dev F1-score	Test F1-score
GPT2	0.948	0.338
Racism RoBERTa	0.818	0.422
Harassment RoBERTa	0.843	0.518
RoBERTa with adapters	0.85	0.659

The results we obtained in the task, compared against the gold test data of the organizers' dataset, The results we obtained in the task, compared against the gold test data of the organizers' dataset, rank sixth in the standings of the ranking for the first subtask of the DETESTS competition, shown in Table 2, with an F1 score of 0.6596.

Table 2
Ranking by teams and obtained F-Score on the first subtask of the DETESTS competition.

Ranking	**Team Name**	**F-Score**
	GoldStandard	1.0000
1	I2C_III	0.7042
2	I2C	0.7005
3	umuteam	0.6990
4	I2C_II	0.6689
5	Lak_NLP	0.6627
6	daminci	0.6596
7	Elias-Urios-Alacreu	0.6438
8	Salsa Version	0.6387
9	MALNIS	0.6382
10	JPG	0.6348

6. Conclusions and Future Work

In this work, we have presented our approaches for the first subtask problem of the DETESTS competition, in which we have envisioned an exclusively Transformer-based approach to the proposed problem. While some could argue that relying on the process of fine-tuning a Transformer-based model could lead to some potential downfalls, we have observed that it produces overall satisfying results in the domain we have worked on, and therefore we decided to adopt this architecture on an exclusive basis for our proposal.

The results of the presented classifier systems have generally been positive and shown good performance on the DETESTS subtask in which we have participated. The results are inspiring if we consider that these systems were trained on meaningful data, manually extracted and carefully selected from authentic user-generated content. With all, the results do not account merely for an exclusively experimental setting, but they also potentially extend to real-world scenarios.

There is room for improvement, however, since some of the limitations in our results could be attributed to a lack of more powerful hardware to train larger-scale models; for instance, it could have been possible to fine-tune larger versions of BERT-based and GPT2-based models, as an experiment on the effects of larger-scale models on the obtained results in the task.

However, the use of larger models would not have necessarily led to better results, and as such, another possible line of work could have been building alternative models while still fundamentally relying on the Transformer architecture. For instance, taking into account the results of the different models we had trained, a new proposal for this project is to concatenate the models we have fine-tuned. With this concatenation, we can use the output layers of a determined model as hidden layers of the holistic model; this means using the output of a model as the input of the next one, allowing some knowledge transfer.

References

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

[2] M. Taulé Delor, A. Ariza, M. Nofre, E. Amigó Cabrera, P. Rosso, Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish, 2021-09.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762. doi:10.48550/ARXIV.1706.03762.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, volume 1, 2019, p. 9.

[5] 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, 2019. URL: https://arxiv.org/abs/1907.11692. doi:10.48550/ARXIV.1907.11692.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-Efficient Transfer Learning for NLP, 2019. URL: https://arxiv.org/abs/1902.00751. doi:10.48550/ARXIV.1902.00751.

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[8] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, 2017. URL: https://arxiv.org/abs/1711.05101. doi:10.48550/ARXIV.1711.05101.

[9] B. Ko, H.-J. Choi, Twice fine-tuning deep neural networks for paraphrase identification, volume 56, 2020, pp. 444–447. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2019.4183. doi:https://doi.org/10.1049/el.2019.4183.

[10] D. Masip, G. Ristow, O. Clavijo, davidmasip/racism, https://huggingface.co/davidmasip/racism, 2022. Accessed: 2022-06-08.

[11] Datathon against Racism, Datathon against Racism, 2022. URL: https://bcnanalytics.com/datathon/.

[12] A. Quizhpe, L. Negrón, D. Pacheco, B. Requenes, P. Pasaca, hackathon-pln-es/Detect-Acoso-Twitter-Es, https://huggingface.co/hackathon-pln-es/Detect-Acoso-Twitter-Es, 2022. Accessed: 2022-06-08.

[13] Hackathon 2022 de PLN en Español, Hackathon 2022 de PLN en Español, https://somosnlp.org/hackathon, 2022. Accessed: 2022-06-08.

[14] 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, volume 68, 2022, pp. 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[15] J. Pfeiffer, A. R"uckl'e, C. Poth, A. Kamath, I. Vuli'c, S. Ruder, K. Cho, I. Gurevych, AdapterHub: A framework for adapting transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 46–54. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.7.

[16] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, 2021.

[17] G. W. Allport, K. Clark, T. Pettigrew, The nature of prejudice, Addison-wesley Reading, MA, 1954.

[18] P. Chiril, F. Benamara, V. Moriceau, "Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification?, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 2833–2844. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.242. doi:10.18653/v1/2021.findings-emnlp.242.

[19] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, B. Y. Zhao, Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–11. URL: https://doi.org/10.1145/3313831.3376488.

[20] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), volume 12, 2018, p. 59.

[21] A. Fokkens, N. Ruigrok, C. Beukeboom, S. Gagestein, W. Van Atteveldt, Studying muslim stereotyping through microportrait extraction, in: H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, T. Tokunaga (Eds.), Proceedings of the LREC 2018, Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2019, pp. 3734–3741. 11th International Conference on Language Resources and Evaluation, LREC 2018 ; Conference date: 07-05-2018 Through 12-05-2018.

[22] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza Morales, J. Gonzalo Arroyo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism Identification in Social neTworks, 2021-09.

[23] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How Do You Speak about

Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants, volume 11, 2021. URL: https://www.mdpi.com/2076-3417/11/8/3610. doi:10.3390/app11083610.

[24] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social Bias Frames: Reasoning about Social and Power Implications of Language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: https://aclanthology.org/2020.acl-main.486. doi:10.18653/v1/2020.acl-main.486.

[25] S. Manuela, C. Gloria, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, CEUR, 2020, pp. 1–9.

[26] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to consider disagreement in evaluation, in: 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, 2021, pp. 15–21.

[27] E. Costa, A. Lorena, A. Carvalho, A. Freitas, A review of performance evaluation measures for hierarchical classifiers, in: Evaluation methods for machine learning II: Papers from the AAAI-2007 workshop, 2007, pp. 1–6.

[28] H. Jain, Y. Prabhu, M. Varma, Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 935–944.