# Is Data Worth More Than Language?

Roberto Labadie Tamayo[1], Reynier Ortega Bueno[2]

[1]Universidad de Oriente, Cuba

[2]PRHLT Research Center, Universitat Politècnica de València, Valencia Spain

**Abstract**

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work. A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work. A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

**Keywords**

Linguistic Ensemble, Paraphrasal Criteria Augmentation, Sexism, Transformers

## 1. Introduction

In the last few years a diversity of social problems has been tackled as Natural Language Processing (NLP) tasks, resulting in the development of robust Artificial Intelligence (AI) models with considerable results. Nevertheless, each of these tasks involves particular challenges taking into account communicative devices employed by humans, which hamper for machines to understand the language determinism.

Sexism is an inherent cultural phenomenon of our society which can be easily verifiable in the way people (independently of their sexual gender) express themselves. Hence social media, as a mirror of the tangible reality, suffer from this phenomenon added to a commonly observed toxicity, with written manifestations ranging from stereotyping and objectification to sexually violent positions of users.

In this biased media, where the data for developing the above-referred models is taken from, it is important to determine whether messages are sexist, to develop unbiased and detoxified software. On the other hand, many studies stating the social networks as a place where besides positive processes converge an important portion of psychological violence, argues for its relation with the emotional and physical health harming of women being target of sexist content [1]. About the concept of "online" and "offline" persona, [2] conclude with certainty

*IberLEF 2022, Spetember 2022, A Coruña, Spain.*

rlabadiet@gmail.com (R. L. Tamayo); rortega@prhlt.upv.es (R. O. Bueno)

CEUR Workshop Proceedings (CEUR-WS.org)

that “interacting with sexist content online can indeed carry over to sexist attitudes offline”. Taking this into account, as long as a platform is able to protect its users from toxicity, it prevents its spreading and helps to stop the trivial assumption of this inherent phenomenon.

Sexism often comes from either offensive or aggressive messages, but sometimes is masked with funny and/or friendly comments; here is where part of the task complexity comes from, since the text isolation of a contextual environment may difficult the sexism detection even for humans. Also, due to the cultural-driven perception of this phenomenon, the availability of annotated data for developing supervised methods of classification is not such a mounting quantity. Finally, sexism perception as well as any other form of verbal language, is subject to additional knowledge from the source, e.g., gestures, prosody features, visual content, and situational environment, which do not accompany textual information in social media.
All of this represent challenging facts for sexism detection task within NLP. Following this idea, the task EXIST: sEXism Identification in Social neTworks at IberLEF 2022 [3], aims at computationally recognizing sexist language on messages from microblogging social media (i.e., Twitter), as well as categorizing it attending to the type of sexism from a multilingual perspective of the phenomenon and as a continuation of the first shared task EXIST@IberLEF 2021 [4].
From last year competition, many models were proposed to face these tasks, the approaches mainly were supported on representations determined from state-of-the-art transformers-based Language Models (LM) [5, 6, 7, 8], specifically the top 10 ranked teams employed these arquitectures. The best performed model proposed by [9], assembled different trasnformer models based on BERT for English and BETO for Spanish and a multilingual variant of BERT combining the prediction of them and taking the higher standardized value in their prediction units. The system proposed by [10] employed a fine-tuned XLM-R over an extended EXIST dataset by incorporanting MeTwo dataset [11] and HatEval 2019 dataset [12]. Other models employed traditional Machine Learning models such as Logistic Regression, Support Vector Machines, Random Forest [13, 14].
Other works have explored individually the detection of stereotyped hate speech spreading taking into account the misogynist form of communication [12, 15] as well as the masked form of this kind of targeted speech with funny messages [16, 17]. In those contexts the systems best ranked also were aligned with the use of sota language models yielding robust models in front of these tasks.
In this working notes we introduce our architecture for participating in the EXIST task at IberLEF 2022 and we study how augmenting the prediction criteria for models ensembles in two different ways influence that kind of approaches. Also we evaluate the expansion of learning examples for our models by introducing data coming from different sources versus expanding them by paraphrasing the provided data in competition. The source code of our approach is available on GitHub [1].

The paper is organized as follows: in Section 2 we briefly introduce a description of the tasks and datasets employed. Section 3 presents the system's architecture and provides details about its modules and methodologies. Section 4 describes the experiments and the achieved results.

[1] https://github.com/labadier/EXIST

Finally, we present our conclusions and provide some directions that we plan to explore in future work.

## 2. Task and Datasets Description

At EXIST 2022 as in 2021 edition, were proposed two subtasks, the first one aiming to determine whether a tweet contains sexist expressions or behaviors. Whereas the second task aimed to determine, in such tweets labeled as sexist, the category of the message according to the type of sexism, i.e., the different facets of women that are undermined: Ideological and inequality, stereotyping and dominance, objectification, sexual violence and mysogyny and no-sexual violence.

Organizers proposed the same training dataset from 2021 edition with roughly 6980 examples, whose data distribution for the first subtask was balanced among positive and negative classes. A different situation with respect to the labeling balance for second subtask is observed, where objectification and sexual-violence categories where softly underrepresented.

We also introduced examples from other datasets where sexism is disguised with humor, or it is employed through an ambivalent way by glorification of traditionally feminine behaviors or demonizing "unladylike" behavior in media coverage with a hateful language.

### 2.1. MAMI Dataset

The task Multimedia Automatic Misogyny Identification (MAMI) [18] consists in the identification of misogynous memes, taking advantage of both text and images available as source of information. From this data we took the text transcription and its annotations for training models in both, first and second subtaks, since the explored categories were almost aligned with the ones evaluated in EXIST. We mapped the annotations from MAMI dataset to EXIST categories as follows; from *shaming* to *ideological-inequality* and *misogyny-non-sexual-violence*, from *stereotype* to *stereotyping-dominance.*

### 2.2. HaHa Dataset

In HAHA@IberLEF2021: Humor Analysis based on Human Annotation [16], was proposed a dateset composed by examples of tweets written in Spanish, annotated regarding the presence of humor, funniness score prediction, the humor mechanism employed (i.e., parody, stereotype, etc.), and the humor target, i.e., for a humorous tweet, the target of the joke from a set of classes such as racist jokes, sexist jokes, etc (what it is making fun of). Here, as examples of sexist tweets, we took all of those labeled as humorous and whose target was related to sexist jokes, afterwards we mapped their annotated targets into their corresponding categories from EXIST task taking into account the mechanism employed (e.g., stereotyping, insults, etc.).

### 2.3. Hahackathon Dataset

For the HaHackathon@SemEval2021: Detecting and Rating Humor and Offense, the dataset provided by the organizers contains English tweets annotated with the presence of humor as well as the humor rating, controversiality and offensiveness rating of the messages. From this dataset, considering the overlapping between humor and offensiveness, we used as positive examples of sexism, and just for subtask 1, those tweets containing popular expressions and terms commonly used to underestimate the role of women in our society and/or spread hate on them and were annotated with values of offensiveness rating greater or equal than 1.0.

## 3. Systems Overview

We explored the performance of two main methods for facing sexism detection and classification tasks, both of them consisting in increasing the data points resulting from mixing the HaHa, Hahackathon, MAMI and EXIST data ($\mathcal{D} - dataset$). In the first one we augmented the points before classification (i.e., in the training stage), via back-translation, taking as pivot-languages different tongs (i). The second strategy relied on augmenting the points at classification stage by assembling the predictions from multiple language models after translating the input message into their respective language (ii).
These methods were based on transformer architectures from the HuggingFace Transformers library[2], from there, we simply finetuned these models over the respective approach dataset taking as target task a multitask perspective involving subtasks 1 and 2. Over this simple paradigm we study their performance.

### 3.1. Data Augmentation

For augmenting data we selected six languages (i.e., es, en, fr, de, pt, it). For approach (i) we paraphrased the tweets translating all the messages into those pivot-languages and then back into Spanish and English, in such a way that for every examples in the original a dataset (i.e., the one composed by all shared tasks data) we introduce 5 new points. For (ii) we keep the points from first step in order to have data for fine-tuning models pretrained in those languages. To carry out, every step the library googletrans[3] was employed.

### 3.2. Transformers Fine-tuning

Transformer models with a RoBERTa-based [7] pre-training procedure over a BERT-base [6] configuration were fine-tuned for each language. Specifically, for English language we used BERTweet [19], for Spanish BETO pretrained on sentiment analysis data [20], for French FlauBERT [21], for Portuguese BERTimbau [22], for Italian the model Italian-BERT [23] and finally for German a BERT model pretrained on German language texts [24].
For all of them we employed as in [25] a gradual unfreezing-discriminative fashion, setting a different learning rate for each encoder layer. Each example was labeled according the annotation obtained in the original dataset aggregation $\mathcal{D} - dataset$, masking the error propagation from units corresponding to subtask 2 in those cases were information regarding to this was not available.

[2] https://huggingface.co/transformers
[3] https://pypi.org/project/googletrans/

### 3.3. Points Prediction Ensembling

As methods for ensembling predictions made by language models we explored (i) a simple majority voting strategy and (ii) making a vector representation of the examples, composed by the softmax layer output of each model and fed it into a parsimonious model. We hypothesized that these expansion of classification criteria, given by the outputs of different models with a pretraining approached over different data domains would introduce a denoising effect into our system.

Following this line, we also explored prediction augmentation, i.e., employing the back-translation technique just at testing time and making predictions for different paraphrases of the input text in order to combine them, since in front textual information with same semantic kernel, language models must preserve their perception of sexist content.

## 4. Experimental Results

In this section we describe the conducted experiments for evaluating the performance of our systems on the training dataset employing a 5-fold cross-validation fashion. For that, we employed accuracy (Acc) metric.

Since both the strategies relied fundamentally in the use of transformer models, at first we tune them adding an intermediate linear layer ReLU-activated between the encoder module and the softmax classification layer optimizing their parameters with the RMSprop algorithm [26] employing as we mention an increasing learning rate from the shallower leayers to the deeper ones with a 0.01 rate from an initial learning rate from the set (1e-5, 2e-5, 3e-5, 5e-5).

Regarding the use or not of external data (*external*) or no (*provided*) over a multitask approach and considering the best combination of 64 units for the intermediate hidden layer, batch size of 64 examples, initial learning rate of 2e-5, the Table 1 shows the results of each language model.

**Table 1**
Employing External Data for each Language Model. acc(task1/task2)

| Strategy | Language | | | | | |
|---|---|---|---|---|---|---|
| | *EN* | *ES* | *DE* | *FR* | *PT* | *IT* |
| external | 0.886/0.756 | 0.887/0.794 | 0.751/0.552 | 0.778/0.585 | 0.773/0.583 | 0.782/0.590 |
| provided | 0.84/0.702 | 0.889/0.691 | 0.701/0.580 | 0.753/0.574 | 0.760/0.551 | 0.413/0.611 |

As we can observe introducing different data from different domains improved the performance in almost all the scenarios. Nevertheless for some languages the models were not able to yield good enough results compared to English and Spanish pretrained LM. We hypothesize that this imbalance in the performance is due to the domain of the data that the models were pretrained on.

We also study if by making the models to learn from both substasks at the same time (*mtl*), they gain enough lingual information to improve the performance or it simply introduce noise to the learning process w.r.t. a single task learning approach (*stl*). Resulting the best strategy for our system combining those both task in just one learning flow as we can se in Table 2.

**Table 2**
Employing External Data for each Language Model. acc(task1/task2)

| Strategy | Language | | | | | |
|---|---|---|---|---|---|---|
| | *EN* | *ES* | *DE* | *FR* | *PT* | *IT* |
| ***mtl*** | 0.886/0.756 | 0.887/0.794 | 0.751/0.552 | 0.778/0.585 | 0.773/0.583 | 0.782/0.590 |
| *stl* | 0.891/0.723 | 0.832/0.690 | 0.738/0.594 | 0.778/0.580 | 0.751/0.524 | 0.656/0.599 |

The evaluation on the test set from EXIST2021 of the first approach from Section 3, i.e., augmenting the data-points at training stage via-backtranslation; shows a high variance with respect to the dev-sets resulting from cross-validating on the training set, as we can see in Table 3. These unbiased results of the models suggest that even when data explored by the models was augmented, their generalization capability was not enough to consider the diverse ways that sexist comments can be expressed.

**Table 3**
Results on EXIST2021 test-set of Augmenting Data-points on Training Stage

| Strategy | Language | | Overall |
|---|---|---|---|
| | *EN* | *ES* | |
| subtask 1 | 0.767 | 0.753 | 0.76 |
| subtask 2 | 0.622 | 0.608 | 0.615 |

For the second approach, i.e., ensembling LMs predictions by majority voting, we tried the different combinations of models predictions, based on the unbalanced individual accuracy observed from Table 2 and Table 1 , without removing, of course, the original languages of the data, i.e., English and Spanish. From these combinations the best performed was the ensemble considering all the languages but German with a clear improvement w.r.t. the first analyzed strategy.

**Table 4**
Results on EXIST2021 test-set of Ensembling Multiple Language Models

| Strategy | Language | | Overall |
|---|---|---|---|
| | *EN* | *ES* | |
| subtask 1 | 0.779 | 0.798 | 0.789 |
| subtask 2 | 0.638 | 0.668 | 0.653 |

As an alternative for this approach, we test the combination of the output from the softmax layer of each of these models which later were employed to train a Support Vector Machines to

avoid the selection of an specific combination of languages. Nevertheless the performance get worst, i.e., an overall accuracy of 0.771 for subtask 1 and 0.591 for subtask 2.
Finally as another way to increasing the data-points at the evaluation stage, we explored back-translation of the input message employing as pivot each of the studied languages, in such a way we again have 5 new contrastive points for major voting one example by employing the same model, i.e., one for English and one for Spanish, resulting again in a worst overall performance of 0.766 for subtask 1 and 0.651 for subtask 2.

Considering these results and in (ii) we simply take the input sequence into different domains where its corresponding model is trained on, and in (i) we paraphrase the input sequence by making the semantic information flowing towards a pivot language and afterwards back to the original language we may hypothesize:

a. For approach (i) some phrases can be contrastive among their different translations in terms of the sexism perception of the evaluated Language Model. This possibly caused when the text flowed from the pivot language back to the original language the performance of the prediction decreased.
b. Taking the texts into different domains and employing specific LM for each domain helped smoothing the noise produced for an individual model in this sexism detection task.

With respect to the best ranked team in the last year competition our approach introduces new data from sexism-related task and extends the use of a wider range of languages, which allow us to outperform their result of 0.780 of acc and makes our system functional not just for english or spanish languages, but also with (fr, de, pt, it). We evaluated denoising the predictions not just by ensembling multiple language models but introducing back-translation on testing-time, which allow us to conclude the existence of vanishing in te sexism perception by the language model when the phrases are taken into another domain where the LM are not trained in.

Regarding official results on this edition of EXIST task we were allowed to make three submissions,the first one employing the majority voting ensemble strategy, achieving an accuracy of 75,80% and and 65,69% for first and second subtasks respectively and as secondary metric evaluated by the organizers 0,7559 and 0.4635 points of macro-average F-measure. As second submission, the denoised predictions obtained from back-translating at testing stage, obtaining an accuraccy of 74,57% and 63,71% for subtasks 1 and 2 respectively whereas for macro-average F-measure 0,7426 and 0,4325 points for each subtask respectively. Finally our worst performed submission considered the softmax output analysis through an SVM, yielding in terms of accuracy 49,05% in subtask 1 and 31,10% in subtask 2, and macro-average F-measures of 0,4872 and 0,1508.

## 5. Conclusion and Future Works

In this paper we describe our system and workflow for facing the task EXIST: sEXism Identification in Social neTworks at IberLEF 2022, consisting of given a message, determining whether it contains sexist undertones as well as the kind of sexism. We addressed this task at the time we study the impact of augmenting the data-points at training stage employing back-translation technique or at testing stage by incorporating different state of the art Languages Models to make individual predictions for a final voting decision when classifying.

We achieved the best performance of our system in competition by combining the prediction of Language Models pretrained on English, Spanish, Portuguese, Italian and French, which were fed with a translated version of the input message into their respective tong reporting an accuraccy of 75,8% and 65,69% for first and second subtasks respectively.

************ hablation wrt the objetive of the paper and results obtained.

## Acknowledgments

## References

[1] S. Berg, C. U. of New York. Ph. D. Program in Social Welfare, Everyday Sexism and Post-traumatic Stress Disorder in Women: A Correlational Study, City University of New York, 2001. URL: https://books.google.com.cu/books?id=XzqvwgEACAAJ.

[2] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline, Comput. Hum. Behav. 52 (2015) 436–442. URL: https://doi.org/10.1016/j.chb.2015.06.024. doi:10.1016/j.chb.2015.06.024.

[3] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022).

[4] F. R.-S. y Jorge Carrillo-de-Albornoz y Laura Plaza y Julio Gonzalo y Paolo Rosso y Miriam Comet y Trinidad Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[6] 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.

[7] 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.

[8] 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.

[9] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, 2021. URL: https://arxiv.org/abs/2111.04551. doi:10.48550/ARXIV.2111.04551.

[10] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepcevic, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic sexism detection with multilingual transformer models ait fhstp@exist2021, in: IberLEF@SEPLN, 2021.

[11] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[12] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://www.aclweb.org/anthology/S19-2007. doi:10.18653/v1/S19-2007.

[13] R. Kumar, S. Pal, R. Pamula, Sexism detection in english and spanish tweets, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 500–505. URL: http://ceur-ws.org/Vol-2943/exist_paper17.pdf.

[14] F.-J. Rodrigo-Ginés, J. C. de Albornoz, L. Plaza, Unedbiasteam at iberlef 2021's exist task: Detecting sexism using bias techniques, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[15] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval (ECIR 2021), Springer, 2021. URL: https://doi.org/10.1007/978-3-030-72240-1_66.

[16] L. C. y Santiago Castro y Santiago Góngora y Aiala Rosa y J. A. Meaney y Rada Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish, Procesamiento del Lenguaje Natural 67 (2021) 257–268. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6394.

[17] J. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval 2021), Association for Computational Linguistics (ACL), 2021, pp. 105–119. URL: https://semeval.github.io/SemEval2021/. doi:10.18653/v1/2021.semeval-1.9, 15th International Workshop on Semantic Evaluation, SemEval 2021 ; Conference date: 05-08-2021 Through 06-08-2021.

[18] E. F. F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-

2022 Task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, 2022.

[19] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[20] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021. arXiv:2106.09462.

[21] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490. URL: https://www.aclweb.org/anthology/2020.lrec-1.302.

[22] F. Souza, R. Nogueira, R. Lotufo, BERTimbau: pretrained BERT models for Brazilian Portuguese, in: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear), 2020.

[23] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[24] O. Guhr, A.-K. Schumann, F. Bahrmann, H. J. Böhme, Training a broad-coverage german sentiment classification model for dialog systems, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1620–1625. URL: https://www.aclweb.org/anthology/2020.lrec-1.202.

[25] R. Labadie Tamayo, D. Castro Castro, R. Ortega Bueno, Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-177.pdf.

[26] G. Hinton, N. Srivastava, K. Swersky, Lecture 6a overview of mini–batch gradient descent, Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online (2012).