FraunhoferSIT@EXIST2024: Leveraging Stacking Ensemble Learning for Sexism Detection

Notebook for the EXIST Lab at CLEF 2024

Shiying Fan[1,*], Raphael Antonius Frick[1] and Martin Steinebach[1]

[1]*Fraunhofer SIT, Rheinstraße 75, Darmstadt, 64295, Germany*

Abstract

The dissemination of inappropriate speeches, such as sexist language, on social media has a negative impact on internet users. To promote the technical development of automatic sexism detection, the EXIST lab has been engaged in this field for the past three years. This paper presents a technical report from the FraunhoferSIT team participating in the EXIST shared task for 2024. To address the issue of detecting sexism in tweets, we have experimented with ensemble learning algorithms. Additionally, we implemented a data augmentation method through synonym replacement using rule-based techniques and language models to increase the size of the training data. We participated in tasks 1 to 3. In general, the proposed system did not demonstrate a competitive performance among other systems in the challenge. However, it was observed that it exhibited a better performance in regression tasks compared to its performance in classification tasks.

Keywords

Sexism Detection, Machine Learning, Stacking Ensemble, Data Augmentation

1. Introduction

The advent of social media platforms and their prevalence have provided individuals with the opportunity to express personal opinions freely and in a timely manner, or to engage in online communication with others. While technology has undoubtedly facilitated interaction between people, it has also contributed to the dissemination of inappropriate content online, such as content with a sexist intention. Since the increased prevalence of discriminatory, harassing, and other forms of sexist content on social media can cause emotional distress to individuals concerned [1], there is an urgent need to enhance techniques for automatic sexism detection on social media to improve the online environment.

Given the variety of linguistic techniques employed in the formation of sexist sentences as stated in [2], the automatic detection of sexism on social platforms remains a challenging task. The EXIST lab, with the objective of advancing the development of methodologies for the automatic detection of sexism on social media, has been engaged in this field for the past three years. The EXIST 2024 shared task builds upon the mission of the previous years and extends the scope of sexism detection from tweets to memes [3, 4].

Previous EXIST shared tasks have involved participants experimenting with different methods and proposing various systems for automatic sexism detection in tweets. Among the methodologies explored, ensemble methods have been widely used. Nevertheless, their application has been mostly in conjunction with language models. Based on the considerable success of machine learning algorithms in performing classification and regression tasks, there is potential for their investigation in the task of sexism detection by leveraging ensemble techniques for machine learning models. Consequently, this paper investigates the utilization of an ensemble of machine learning models with the help of stacking to enhance the accuracy and robustness of machine learning system for sexism detection. Our paper contributes to the exploration and development of techniques for the research field of the automatic sexism detection on social media.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

shiying.fan@sit.fraunhofer.de (S. Fan)

The structure of the paper is organized as follows: firstly, we present the application of ensemble learning for automatic sexism detection in the past EXIST challenge. Then, we provide a description of the tasks performed and the dataset we used for these tasks. After that, we depict the methodology employed to address the given tasks and outline our experimental setup. Subsequently, we report on the results we achieved in the competition and discuss the limitations of the proposed system. Finally, we conclude this paper with an overview of future work.

2. Related Work

A review of the EXIST competition over the past three years found that the potential of various learning methods suitable for NLP tasks has been widely discussed by participants. The techniques for the detection of sexism that have been explored include contrastive learning, transfer learning, ensemble learning and others. While both Transformer models [5] and traditional machine learning models such as Support Vector Machines or Naive Bayes classifiers have been employed for the implementation of various techniques, Transformer-based models continue to be the predominant choice.

The Transformer-based language models have been specifically designed to address NLP tasks and achieve state-of-the-art performance in numerous NLP benchmarks. In previous challenges, both monolingual [6] and multilingual [7] Transformer models have been utilized for the purpose of detecting instances of sexism. Additionally, [8] compared the performance of monolingual and multilingual models.

Despite the significant advances in NLP tasks achieved by language models, the performance of individual models in classification tasks remains limited. Consequently, ensemble methods are frequently employed to enhance the capabilities of individual models. For instance, [9] and [10] explored the potential of combining multilingual Transformer models and ensemble learning techniques for providing the final prediction. In contrast, [11], [12] and [13] investigated the ensembles of monolingual models for sexism detection. Moreover, [14] examined and compared the classification performance of several multilingual models and monolingual models. The ensemble combinations of language models they selected were not based on the training type of model regarding their linguistic capability. Instead, they were combined with an English model, a Spanish model, and an additional baseline model. Further examples of the investigation of ensembles of Transformer models can be found in [15, 16, 17].

Although ensemble approaches have been demonstrated to be effective in performing the given shared task in comparison to the performance of baseline models, [18] proposed a novel approach that combines transfer learning and ensemble learning. This approach was employed to enhance the feature learning effect of pre-trained and fine-tuned Transformer models. Furthermore, [19] proposed a bi-ensemble method, which merges two ensemble approaches: an ensemble consisting of two architectures and an ensemble composed of several trained models of the same architecture.

In addition to the study of ensembles of language models, ensembles of machine learning models have also been examined. For instance, the approach taken by [20] involved the use of an ensemble of machine learning models, whereby the probabilities predicted by these models were aggregated. Nevertheless, despite their potential, the ensembles of machine learning models have not been subjected to sufficient investigation within the context of the EXIST challenge. Hence, it is necessary to conduct further research into their capabilities.

3. Task Description

The EXIST 2024 shared task comprises two parts: the classification of tweets (Tasks 1 to 3) and the classification of memes (Tasks 4 to 6) [3]. Our participation was limited to the classification of tweets. Among the tweet classification tasks, Task 1 is formulated as a binary classification task, in which the intention of sexism is identified in tweets. Each tweet is expected to be classified into one of the two categories: either *yes* or *no*. Task 2 is formulated as a multi-class classification task. Tweets identified as sexist will be further categorized into one of the three categories based on the intention of the source,

namely, *direct*, *reported*, or *judgmental*. Task 3 is formulated as a multi-label classification task. Tweets predicted to be sexist are further categorized into one or more of the following categories: *ideological-inequality*, *stereotyping-dominance*, *objectification*, *sexual-violence*, and/or *misogyny-non-sexual-violence*.

4. Dataset

The process of identifying instances of sexism in texts is inherently subjective. To enhance the learning effect of the system and avoid author bias, EXIST 2024 employs a Learning With Disagreement (LeWiDi) paradigm for the development of the dataset and the evaluation of the systems [3]. This enables the system to consider the perspectives of multiple annotators, thereby facilitating a more equitable learning process. In accordance with the LeWiDi paradigm, each tweet was annotated by six annotators of different genders, ages, ethnicities, educational backgrounds, and countries.

Table 1
EXIST 2024 Tweets Dataset

	English Tweets	Spanish Tweets	Total
Train Set	3260	3660	6920
Dev Set	489	549	1038
Test Set	978	1098	2076

Since we were only engaged in the tweet classification tasks, we utilized only the tweets dataset provided by the EXIST shared task. As shown in Table 1, the tweets dataset is divided into three distinct sets: the train set, the development set, and the test set. Both the train set and the development set comprise labeled tweets in English and Spanish, with a total of 6920 and 1038 tweets, respectively. The test set, on the other hand, includes 2076 unlabeled English and Spanish tweets.

5. Methodology

As previously stated (see Section 3), our team participated in the first three tasks related to tweet classification. These tasks can be identified as three types of classification problems: binary classification (Task 1), multi-class classification (Task 2), and multi-label classification (Task 3). In addition to providing hard labels for tweet classification, we also submitted soft labels for each task, which are considered to target the issue of regression. In this section, we will present the methodology employed in the given shared task to address classification and regression problems.

5.1. Preprocessing

Tweets often contain hashtags, mentions, and URLs. The removal of mentions and URLs from tweets has been demonstrated to have a minimal effect on the interpretation of the original tweets. Hashtags typically represent the topics of tweets. Since the topics of tweets have been predefined for the purpose of sexism detection, the hashtags are considered to have limited importance. Therefore, in the preprocessing stage, hashtags, mentions, and URLs were removed from the tweets properly.

Emojis are frequently utilized in tweets as a means of expressing emotions. They are typically employed to convey the attitude of the author, to extend and strengthen the emotion expressed, or to reverse the textual meaning in instances of sarcasm or irony. Given the potential for emojis to influence the interpretation of tweets, it is recommended that they are treated as a constituent of tweet texts. Consequently, all emojis were retained and converted into text format using the emoji python library [21]. Emojis from Spanish tweets were converted into Spanish text, while emojis from English tweets were converted into English text.

5.2. Data Augmentation

Data augmentation encompasses a range of techniques designed to increase the quantity and diversity of the training data associated with a given dataset, thereby enhancing the accuracy and robustness of machine learning systems [22]. It is used to address the issue of overfitting on the training set, which can arise due to limited labeled data [23]. Furthermore, it serves to mitigate bias and fix class imbalance [24]. Data augmentation, which originated from research fields of computer vision [25], has attracted considerable attention in the NLP community in recent years. The data augmentation methods specialized for NLP tasks have been actively investigated to address the issues of limited data for specific NLP research areas.

Despite the advantages that data augmentation techniques bring to text classification settings, data augmentation remains a challenging task due to the difficulty of defining textual transformations that preserve the labels. To ensure the compatibility of the labels on the original texts with those on the augmented texts, we employed two augmentation methods: synonym replacement [26] and contextual augmentation [27]. Both of these methods were applied at the word level. These approaches have been demonstrated to maintain the labels of the original texts, as the semantic meanings of the augmented versions remain unaltered.

WordNet [28, 29] is an English lexical database that groups words into sets of synonyms. The implementation of the synonym replacement method using WordNet for English tweets has been demonstrated to provide high-quality data. Due to the limitations of the available Spanish synonym lexical database, we selected three different Transformer language models from HuggingFace [30] for the implementation of the contextual augmentation method. The models employed include BERTIN [31], ALBERT Base Spanish [32], and RoBERTuito [33]. In total, the original train and development datasets were subjected to a tenfold augmentation using the aforementioned methods.

5.3. Stacking Ensemble Method

Ensemble methods are learning algorithms that are constructed based on the combinations of a set of learning algorithms with the objective of enhancing the overall prediction performance of multiple single learning algorithms. The rationale behind the development of these ensembles is that the error rate induced by the ensemble of different learning algorithms has the potential to be compensated in comparison to the prediction performance of single learning algorithms [34, 35]. In addition, the accuracy and diversity of the individual learning algorithms employed in the construction of the ensemble still play a critical role in enhancing the robustness of the ensemble [36].

A variety of techniques has been developed for constructing ensembles, including bagging (bootstrap aggregating) [37], boosting [38, 39], AdaBoost (adaptive boosting) [40], voting [41], and stacking [42]. As stated in [35], ensemble methods are considered the state-of-the-art solution for many machine learning challenges. They have also been widely used in previous years for the EXIST shared task, as outlined in Section 2. However, the ensemble learning methods reported in the EXIST working notes are predominantly in conjunction with Transformer models. The ensemble of machine learning models remain under-researched for the detection of sexism in tweets. Thus, for the EXIST 2024 shared task, we utilize ensemble of machine learning models.

In the presented tasks, both soft labels and hard labels can be provided for tweets. The task of providing soft labels for tweets can be considered a problem of regression, while the task of providing hard labels for tweets is referred to as a classification problem. As the ensemble method of stacking machine learning models can be applied to both classification and regression problems, we employed the stacking ensemble to address the tasks outlined in Section 3. In the context of resolving these two problems, we differentiate in our approach to model selection for the stacking ensemble. All models employed were implemented from the Scikit-learn library [43].

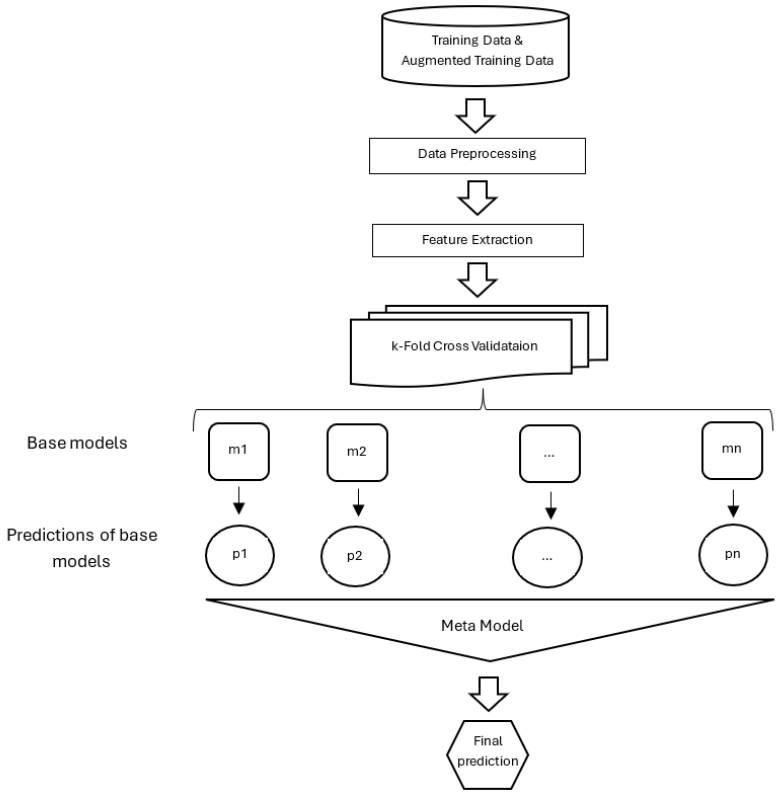

Figure 1: Experimental setup

6. Experimental Setup

Figure 1 illustrates the experimental setup we employed in the EXIST challenge. As depicted in the figure, we initially addressed the data issue for both classification and regression tasks. We applied the augmentation methods described in Section 5.2 to expand the size of the given dataset to ten folds. Thereafter, we preprocessed the tweets in accordance with the approach proposed in Section 5.1. Then, we extracted the textual features of the clean tweets using the Bag-of-Words method. Subsequently, we employed a five-fold cross validation procedure to partition the provided data samples. The data samples were then subjected to training using multiple distinct base models. Finally, the predictions of the base models were incorporated into a meta-estimator for the purpose of making the final prediction.

In order to perform the classification tasks, we selected five different models as base models. The selected models included Multinomial Naive Bayes (MNB), Stochastic Gradient Descent (SGD), Decision Tree (DT), k-Nearest Neighbors (kNN), and Logistic Regression (LR). With regard to the meta-model, we chose Extra Trees (ET).

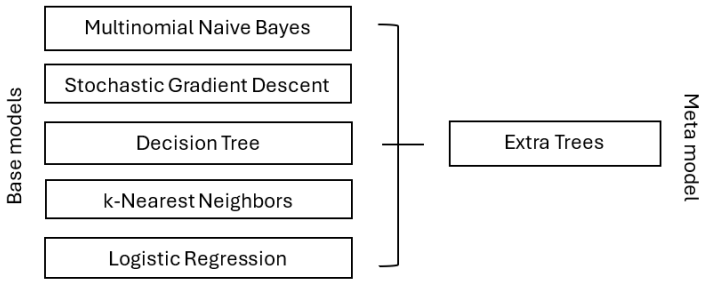

Figure 2: Stacking classifiers

The base models selected for the given classification tasks are commonly employed in supervised learning and can be applied to both classification and regression problems. Nevertheless, it has been demonstrated that MNB, DT, kNN, and LR are more suitable for classification tasks than for regression tasks. In contrast, SGD presents an optimisation technique that can be used to optimise models for both classification and regression tasks. Additionally, all of these models demonstrate scalability to large datasets and offers various methods to manage overfitting.

The ET, also known as Extremely Randomised Trees, is a tree-based ensemble method for supervised classification and regression problems. Its distinctive nature involves randomizing both attribute and cut-point choices while splitting a tree node and building completely randomized trees whose structures are independent of the output values of the learning sample [44]. This characteristic suggests that it has the potential to be used as a meta-estimator for stacking ensemble learning.

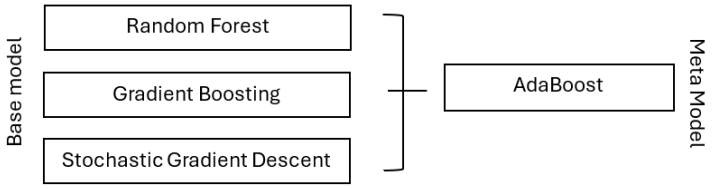

Figure 3: Stacking regressors

In the case of the regression tasks, we applied Random Forest (RF), Gradient Boosting (GB), and Stochastic Gradient Descent (SGD) as base models, and AdaBoost as the meta-model. In essence, despite the differences in their methodologies and types of base learners, these algorithms share the common objective of enhancing prediction accuracy. This can be attained by employing either a tree-based ensemble learning approach, comprising RF, GB, and AdaBoost, or an optimisation technique, such as SGD. Furthermore, each of these algorithms has been developed to handle issues such as overfitting, scalability, and the inherent bias-variance trade-off in the learning process.

7. Results and Discussion

In this section, we present the official evaluation results of our submissions to the competition (see Table 2, 3, and 4). The evaluations were performed in two modes: Hard-Hard evaluation and Soft-Soft evaluation, using the official metric Information Contrast Measure (ICM) and selectively the F_1 score [4]. Additionally, each of these two modes was used for evaluating English and Spanish tweets, respectively.

Table 2
Evaluation results for Task 1

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Soft-Soft ALL	24/40	−0.0658	0.4895	0.8801
Soft-Soft ES	26/40	−0.0344	0.4945	0.8475
Soft-Soft EN	25/40	−0.1593	0.4744	0.9167
	Rank	**ICM-Hard**	**ICM-Hard Norm**	F_1
Hard-Hard ALL	53/70	0.2320	0.6166	0.6823
Hard-Hard ES	53/66	0.2009	0.6005	0.7062
Hard-Hard EN	55/68	0.2334	0.6191	0.6447

In the first task of identifying instances of sexism in tweets, we achieved a general ICM score of 0.4895 for the Soft-Soft evaluation and 0.6166 for the Hard-Hard evaluation. In the second task of categorising the source intention in tweets, we achieved a general ICM score of 0.1708 for the Soft-Soft evaluation and 0.3665 for the Hard-Hard evaluation. In the third task of sexism categorisation, the

Table 3
Evaluation results for Task 2

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Soft-Soft ALL	21/35	−4.0856	0.1708	1.7649
Soft-Soft ES	22/35	−4.1019	0.1715	1.7784
Soft-Soft EN	21/35	−4.0500	0.1690	1.7498
	Rank	ICM-Hard	ICM-Hard Norm	Marco F_1
Hard-Hard ALL	34/46	−0.4106	0.3665	0.3832
Hard-Hard ES	34/46	−0.3871	0.3791	0.4078
Hard-Hard EN	35/46	−0.4801	0.3339	0.3367

Table 4
Evaluation results for Task 3

	Rank	ICM-Soft	ICM-Soft Norm	
Soft-Soft ALL	14/33	−5.1905	0.2259	
Soft-Soft ES	14/33	−5.3003	0.2241	
Soft-Soft EN	13/33	−4.9748	0.2274	
	Rank	ICM-Hard	ICM-Hard Norm	Macro F_1
Hard-Hard ALL	21/34	−0.7437	0.3273	0.3724
Hard-Hard ES	23/34	−0.6734	0.3496	0.4044
Hard-Hard EN	24/34	−0.8394	0.2943	0.3256

general ICM score was 0.2259 for the Soft-Soft evaluation and 0.3273 for the Hard-Hard evaluation. The results achieved in the competition indicate that both the classification and regression systems proposed performed the best in Task 3 in comparison to their performance in Tasks 1 and 2. This was particularly evident in the Soft-Soft evaluation.

Although the stacked ensemble of machine learning models employed were tailored to address the classification and regression tasks for the EXIST challenge, they did not demonstrate a competitive performance among other systems in the challenge. Given the absence of a universal methodology for the selection of the best model combinations for stacking ensemble, our approach involved the identification of models based on their individual performance on the tasks of sexism detection. Furthermore, the time-consuming nature of the stacking ensemble method precluded the development of the optimal combination of machine learning models to target the characteristic of each task.

8. Conclusion and Future Study

In this paper, we presented our participation in the EXIST 2024 shared task [3, 4]. We introduced the tasks we participated in and the dataset we used for the tasks. Furthermore, we described the methodology employed for sexism detection and reported on the results achieved in the challenge. We also discussed issues that may influence the results.

The official evaluation results indicate that the system proposed in this report did not perform competitively as other systems within the challenge. Nevertheless, the system exhibited comparatively better performance in regression tasks than in classification tasks.

For future work, it would be beneficial to experiment with various ensemble combinations and to gain a more in-depth insight into model choice in order to identify the most suitable ensemble of models for sexist context. In addition to machine learning models, Transformer models can also be stacked to outperform their single performance on both classification and regression tasks.

References

[1] S. Pingree, R. P. Hawkins, M. Butler, W. Paisley, A scale for sexsim, Journal of Communication 26 (1976) 193–200.

[2] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 - learning with disagreement for sexism identification and characterization (extended overview), in: CLEF 2023: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-070.pdf.

[3] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[4] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[6] C. Jhakal, K. Singal, M. Suri, D. Chaudhary, B. Kumar, I. Gorton, Detection of sexism on social media with multiple simple transformers, in: CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-082.pdf.

[7] J. A. García-Díaz, R. Pan, R. Valencia-García, Umuteam at exist 2023: Sexism identification and categorizsation fine-tuning multilingual large language models, in: CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-080.pdf.

[8] K. Bengoetxea, A. Aguirregoitia, Multiaztertest@exist-iberlef2022: Sexism identification in social networks, in: IberLEF 2022, CEUR Workshop Proceedings, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper8.pdf.

[9] P. Cordón, J. Mata, V. Pachón, J. L. Domínguez, I2c-uhu at clef-2023 exist task: Leveraging ensembling language models to detect multilingual sexism in social media, in: CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-076.pdf.

[10] A. F. Magnossão de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023– sexism characterization using large language models under the learning with disagreements regime, in: CLEF 2023: Conference and Labs of the Evaluation Forum, volume 3497, 2023. URL: https://ceur-ws.org/Vol-3497/paper-084.pdf.

[11] A. F. Magnossão de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, in: IberLEF 2021, Málaga, Spain, 2021. URL: https://ceur-ws.org/Vol-2943/exist_paper2.pdf.

[12] L. Davies, M. Baldracchi, C. A. Borella, K. Perifanos, Transformer ensembles for sexism detection, in: IberLEF 2021, Málaga, Spain, 2021. URL: https://ceur-ws.org/Vol-2943/exist_paper5.pdf.

[13] A. Vaca-Serrano, Detecting and classifying sexism by ensembling transformers models, in: IberLEF 2022, CEUR Workshop Proceedings, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper3.pdf.

[14] V. P. Álvarez, J. M. Vázquez, W. Chibane, J. L. D. Olmedo, Automatic sexism identification using an ensemble of pretrained transformers, in: IberLEF 2022, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper6.pdf.

[15] A. F. Magnossão de Paula, R. F. da Silva, Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models, in: IberLEF 2022, A Coruña, Spain,

2022. URL: https://ceur-ws.org/Vol-3202/exist-paper2.pdf.
[16] J. A. García-Díaz, S. M. Jiménez-Zafra, R. Colomo-Palacios, R. Valencia-García, Umuteam at exist 2022: Knowledge integration and ensemble learning for multilingual sexism identification and categorization using linguistic features and transformers, in: IberLEF 2022, CEUR Workshop Proceedings, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper14.pdf.
[17] A. Younus, M. A. Qureshi, A framework for sexism detection on social media via byt5 and tabnet, in: IberLEF 2022, CEUR Workshop Proceedings, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper15.pdf.
[18] J. Böck, M. Schütz, D. Liakhovets, Q. Satriani, A. Babic, D. Slijepčević, M. Zeppelzauer, A. Schindler, Ait_fhstp at exist 2023 benchmark: Sexism detection by transfer learning, sentiment and toxicity embeddings and hand-crafted features, in: CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-074.pdf.
[19] E. Villa-Cueva, F. Sanches-Vega, A. P. López-Monroy, Bi-ensemble of transformer for online bilingual sexism detection, in: IberLEF 2022, CEUR Workshop Proceedings, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper4.pdf.
[20] S. Ravi, S. Kelkar, A. K. Madasamy, Lstm-attention architecture for online bilingual sexism detection, in: CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2023. URL: https://ceur-ws.org/Vol-3497/paper-089.pdf.
[21] T. Kim, K. Wurster, Emoji terminal output for python, 2024. URL: https://github.com/carpedm20/emoji/.
[22] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: IEEE Conference on Computer Vision and Pattern Recognistion, 2019, pp. 113–123. URL: https://doi.org/10.1109/CVPR.2019.00020.
[23] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An empirical survey of data augmentation for limited data learning in nlp, in: Transactions of the Association for Computational Linguistics, volume 11, 2023, pp. 191–211. URL: https://doi.org/10.1162/tacl_a_00542.
[24] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, 2021. URL: https://doi.org/10.48550/arXiv.2105.03075.
[25] M. Bayer, M.-A. Kaufhold, C. Reuter, A survey on data augmentation for text classification, in: ACM Computing Surveys, volume 55, 2022, pp. 1–39. URL: https://doi.org/10.1145/3544558.
[26] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Langauge Processing, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670.pdf.
[27] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, in: Proceedings of NAACL-HLT, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 452–457.
[28] G. A. Miller, Wordnet: A lexical database for english, Communications of the ACM 38 (1995) 39–41.
[29] C. Fellbaum, WordNet: An Electronic Lexical Database, MA: MIT Press, Cambridge, 1998.
[30] 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, Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.
[31] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.
[32] J. Cañete, Albert base spanish, HuggingFace, 2022. URL: https://huggingface.co/dccuchile/albert-base-spanish.
[33] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language

model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[34] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–15. URL: https://doi.org/10.1007/3-540-45014-9_1.

[35] O. Sagi, L. Rokach, Ensemble learning: A survey, WIREs Data Mining Knowl Discov 8 (2017) e1249. URL: https://doi.org/10.1002/widm.1249.

[36] L. K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 993–1001. doi:10.1109/34.58871.

[37] L. Breiman, Bagging predictions, Machine Learning 24 (1996) 123–140. URL: https://link.springer.com/article/10.1007/BF00058655.

[38] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2001) 1189–1232.

[39] J. H. Friedman, Stochastic gradient boosting, Computational Statistics & Data Analysis 38 (2002) 367–378.

[40] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Science 55 (1997) 119–139.

[41] M. Panda, M. R. Patra, Ensemble voting system for anomaly based network intrusion detection, International Journal of Recent Trends in Engineering 2 (2009).

[42] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[44] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (2006) 3–42. doi:10.1007/s10994-006-6226-1.