NLP-LTU at SemEval-2023 Task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset

Sana Al-Azzawi György Kovács Filip Nilsson Tosin Adewumi
Marcus Liwicki
EISLAB Machine Learning, Luleå University of Technology, 977 54 Luleå, Sweden
firstname.lastname@ltu.se

Abstract

In this paper, we propose a methodology for task 10 of SemEval23, focusing on detecting and classifying online sexism in social media posts. The task is tackling a serious issue, as detecting harmful content on social media platforms is crucial for mitigating the harm of these posts on users. Our solution for this task is based on an ensemble of fine-tuned transformer-based models (BERTweet, RoBERTa, and DeBERTa). To alleviate problems related to class imbalance, and to improve the generalization capability of our model, we also experiment with data augmentation and semi-supervised learning. In particular, for data augmentation, we use back-translation, either on all classes, or on the underrepresented classes only. We analyze the impact of these strategies on the overall performance of the pipeline through extensive experiments. while for semi-supervised learning, we found that with a substantial amount of unlabelled, in-domain data available, semi-supervised learning can enhance the performance of certain models. Our proposed method (for which the source code is available on Github[1][2]) attains an $F1$-score of 0.8613 for sub-taskA, which ranked us 10th in the competition.

1 Introduction

Remarkable technological advancements have made it simpler for people from diverse backgrounds to interact through social media using posts and comments written in natural language. These opportunities, however, come with their own challenges. Hateful content on the Internet increased to such levels that manual moderation cannot possibly deal with it (Gongane et al., 2022). Thus, precise identification of harmful content on social media is vital for ensuring that such content can be discovered and dealt with, minimizing the risk of victim harm and making online platforms safer and more inclusive.

Detecting online sexism on social media remains a challenge in natural language processing (NLP), and the Explainable Detection of Online Sexism (EDOS) shared task on SemEval23 (Kirk et al., 2023) addresses this problem. The task has three main sub-tasks: (i) task A; binary sexism detection, in which we determine whether a given sentence contains sexist content, (ii) task B; sexism classification, which places sexist sentences into four categories: threats, derogation, animosity, and prejudiced discussions, and (iii) task C; fine-grained vector of sexism, an eleven-class categorization for sexist posts in which systems must predict one of 11 fine-grained vectors.

One major challenge of this task is the imbalanced class distribution. For instance, sub-task A consists of only 3398 sexist posts, and 10602 non-sexist ones. Using an imbalanced dataset to train models can result in prediction bias towards the majority class (Johnson and Khoshgoftaar, 2019).

In this paper, we (team NLP-LTU) present the automatic sexism detection system developed and submitted to SemEval23 task 10; EDOS. The objective of this study is (i) to examine how different state-of-the-art pre-trained language models (PLM) perform in sexism detection and classification tasks, and (ii) to contribute towards answering the following research question (RQ): **To what extent can data augmentation improve the results and address the data imbalance problem?**

The core of our approach is a voting-based ensemble model consisting of three pre-trained language models: BERTweet-large (Nguyen et al., 2020), DeBERTa-v3-large (He et al., 2021), and RoBERTa-large (Liu et al., 2019). Additionally, in order to address the issue of data imbalance and to expand our dataset, our system's pipeline employed techniques such as data augmentation and semi-supervised learning. We achieved competi-

[1] github.com/SanaNGU/semeval23-task10-sexism-detection-
[2] huggingface.co/NLP-LTU/bertweet-large-sexism-detector

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1421–1427

tive results, ranking us in the top ten for Task A.[3] Our results suggest that (i) using PLMs trained on domain-specific data (e.g. BERTweet-large) leads to better results than using PLMs pre-trained on other sources (ii) In most cases extending all classes via augmentation leads to higher classification scores than using augmentation on the minority classes only to completely balance the class distribution. However, drawing conclusive inferences would require further experiments with multiple data augmentation methods and datasets. (iii) with a substantial amount of unlabelled, in-domain data available, semi-supervised learning can enhance the performance of certain models.

The rest of the paper is organised as follows: in Section 2, we present prior related work; in Section 3, we discuss the proposed system. Then, we describe the experiments in Section 4. Section 5, presents results and error analysis. Finally, we conclude the work in Section 6 and describe what further has to be done.

2 Related Work

In the following section we discuss already existing efforts on the detection of sexism, and efforts directed at data augmentation.

2.1 Sexism Detection

Detecting sexism in social media is essential to ensure a safe online environment and to prevent the negative impact of being a target of sexism. Therefore, several studies have developed datasets and machine-learning models to identify and detect sexism in social media (Nilsson et al., 2022). Waseem and Hovy's early study involves collecting 16K English tweets and annotating them into three categories: racism, sexism, and neutral (Waseem and Hovy, 2016). Similarly, but from a multilingual perspective, Rodríguez-Sánchez et al. (2020) created the MeTwo dataset to identify various forms of sexism in Spanish Tweets, and they use machine learning techniques, including both classical and deep learning approaches. Several additional datasets have since been created to examine a wide range of sexist statements (Parikh et al., 2019; Samory et al., 2021; Rodríguez-Sánchez et al., 2021, 2022).

The aforementioned studies often categorize sexist content into a limited number of classes, typically two to five, without any further breakdown. However, sexist sentences/posts should be identified, and the reasons for the identification should be provided to increase the interpretability, confidence, and comprehension of the judgments made by the detection system. The EDOS (Kirk et al., 2023) task aims to target this problem with fine-grained classifications for sexist content from social media.

2.2 Data Augmentation

A dataset may have several shortcomings that make text classification difficult. This paper mainly focuses on using data augmentation to deal with class imbalance. Easy Data Augmentation (EDA) (Wei and Zou, 2019) use four simple word-based operations to generate new data: synonym replacement, random insertion, random swap, and random deletion. EDA shows that the classification performance improves even with a simple data augmentation approach. Similarly, Kobayashi (2018) stochastically replaces words in the sentences with other relevant words using bidirectional recurrent neural networks.

In more recent studies, PLM are used to get text samples that are more diverse and linguistically correct. Anaby-Tavor et al. (2020) apply GPT-2 to generate synthetic data for a given class in text classification tasks. Another study by Sabry et al. (2022), uses conversational model checkpoint created by Adewumi et al. (2022).

3 System Overview

This section outlines the system pipeline employed in our study, as depicted in Figure 1. The proposed approach entails two main stages, generating additional training samples (Module 1), and classification (Module 2). Each is described in its own subsection below.

3.1 Module 1.A: Data Augmentation

Imbalanced data might impede a model's ability to distinguish between highly-represented classes (e.g., non-sexist) and under-represented ones (i.e., sexist). To address this concern, we studied the potential influence of data augmentation approaches on the system's performance.

We expand module 1.A of Figure 1 to describe the data augmentation module (shown in Figure 2). The module comprises three main steps. First, we fine-tune our best-performing model; (BERTweet-large) using the gold-labelled data. Then, each sentence undergoes two rounds of back translation

[3] https://github.com/rewire-online/edos/blob/main/leaderboard

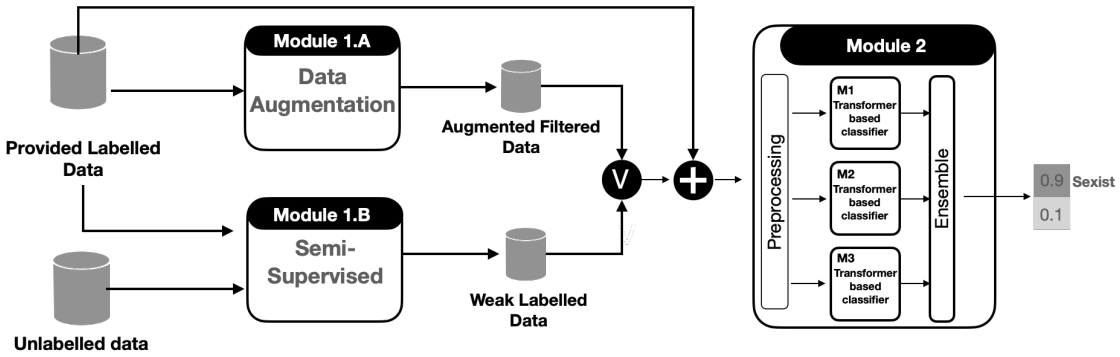

Figure 1: Architecture of the proposed approach

(English to German, and back to English, then English to Russian, and back to English again). Here, our choice of data augmentation method was motivated by its simplicity and the fact that it does not rely on specific task data and It can be applied independently of the task at hand (Longpre et al., 2020).

In the final step, the newly generated English sentences from each stage in the second step are filtered using the fine-tuned model from step one. This ensures that each new synthetic sentence retains its original label. This technique can be employed in two ways. Firstly, it can augment only the underrepresented class (sexist sentences to balance the dataset. Alternatively, both classes can be augmented to double the dataset. We investigate the performance of the data augmentation technique using both ways.

3.2 Module 1.B: Semi-supervised Learning

Two more unlabelled datasets, each with one million entries, were made available by the task's organizers. Inspired by earlier research (e.g. (Shams, 2014)), we used the provided unlabelled datasets to generate weakly labelled samples to balance the original dataset.

As shown in Figure 3, Module 1.B comprises three stages. The first stage being fine-tuning a select pre-trained model (BERTweet-large), using the gold labels. Then, we use the resulting model to create weak labels for the unlabelled data. Lastly, we select samples labelled with a minority class, where the predicted probability of the weak label is at least 0.9.

3.3 Module 2: Ensemble

Similar to the full pipeline, Module 2 can also be broken down into its individual constituents, which are (i) the pre-processing module, (ii) the individual classifiers, and (iii) the ensembling method to combine the decision of these classifiers. Firstly, a pre-processing step is needed, as the data for the tasks was collected from noisy resources (Reddit, and Gab). For this, we used the same common techniques for all models. In particular, we converted all uppercase characters to lowercase, removed repetitive patterns like "heeeey" and additional spaces, eliminated special characters like emojis and hashtags (#), and deleted numbers.

For the individual classifiers, we examined different PLMs such as BERT (Devlin et al., 2019), RoBERTa , and DeBERTa. each of these models

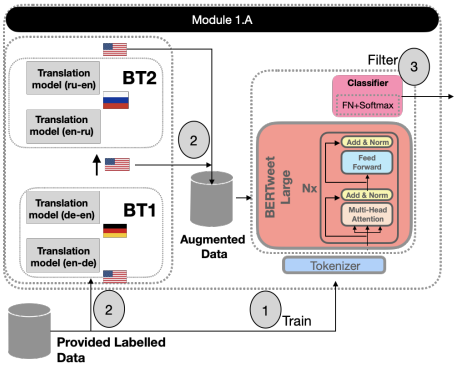

Figure 2: Back translation data augmentation Block

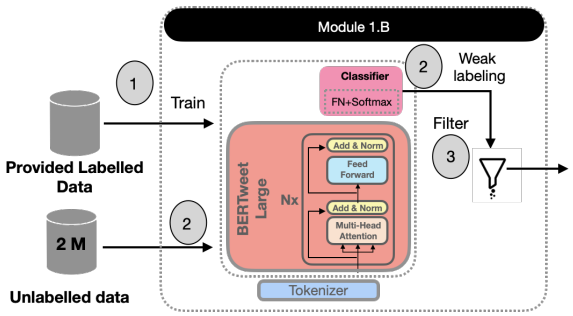

Figure 3: Semi-Supervised Block.

was initially fine-tuned using the entire dataset.

Lastly, we employed an ensemble of the three best-performing classifiers from the previous step for the final submission, namely BERTweet-large, DeBERTa-v3-large, and RoBERTa-large. Ensembling multiple models can potentially prevent egregious mistakes made by a single model (Ruta and Gabrys, 2005; Zhang et al., 2014).

We used two ensemble approaches: majority voting, a hard voting method where the prediction of each classifier is treated as a vote, and the class with the most votes is ultimately selected as the predicted class, and soft average ensemble, in which the output of each model is averaged as shown in Equation 1.

$$y_{final} = argmax(\frac{y_1 + y_2 + y_3}{3}) \quad (1)$$

4 Experimental Setup

All experiments have been implemented using the PyTorch and HuggingFace libraries (Wolf et al., 2020) on an 8 32GB Nvidia V100 GPU-equipped DGX-1 cluster. The server contains 80 CPU cores with the Ubuntu 18 operating system. When evaluating solutions, the macro-averaged $F1$-score was the primary metric.

4.1 Task A : Binary Sexism Detection

In Task A, we employed the proposed pipeline illustrated in Figure 1. Initially, we utilized module 1.A to augment the sexist samples, thereby achieving dataset balance. Subsequently, we integrated the synthetic data with the original data and fine-tuned various pre-trained language models.

The batch size is set to 16, and the Adamw optimizer is used for training. We set the learning rate of the pre-trained model for each language model to 1e-5 and fine-tuned it for three epochs. In the semi-supervised learning context, we utilized identical parameters and generated 7,000 additional samples with sexist content to balance the dataset.

4.2 Task B:sexism classification

For task B, we excluded Module 1.B. from our pipeline, as in our initial experiments, we were not able to train a sufficiently reliable classifier for the weak-labelling on this smaller dataset. Our hyperparameters for this task were the same as discussed above, with the exception of an increased number of epochs (4) used here.

About task B, we exclusively employed Module 2 and Module 1.A. Our rationale for this choice stemmed from the inadequacy of the training dataset, which rendered it unfeasible to produce weak labels for this particular task.

4.3 Task C : Fine-grained Vector of Sexism

In our experiments for task C, due to limited time, we forewent the first modules, and focused on Module 2, fine-tuning several pre-trained language models. We have, however, only used these models individually, as the fine-tuned models did not attain comparable levels of performance to those achieved in the previous tasks. Furthermore, the use of an ensemble in such cases may potentially detract from the overall performance of the system.

5 Results

5.1 Evaluation Phase

During the evaluation phase, we used the development set provided by the organizers. The results for task A are shown in Table 1. Concerning the data augmentation component, we compared two distinct data augmentation strategies. The initial approach entailed doubling the size of the entire dataset, while the alternative strategy solely augmented samples that contained sexist content.

Table 2 shows the results on the development set for Task B. Due to the limited size of the training set, the use of data augmentation techniques resulted in an improved performance for some models, while others exhibited similar $F1$-scores to those obtained without augmentation.

The results shown in Table 1 indicate that the use of the provided dataset with a hard ensemble strategy yields the best performance. Furthermore, the semi-supervised approach improves the performance of some pre-trained models (e.g. BERT-base, HateBERT (Caselli et al., 2020), BERTweet-base), but not those models, which had been pre-trained on larger datasets (e.g. DeBERTa-large-v3, BERTweet-large). We hypothesize that these larger

Model	w/o DA	with DA-double	with DA-balanced	semi-supervised double-	semi-supervised-balanced
BERT-base	82.00	81.5	78.00	81.79	82.10
RoBERTa	83.00	83.5	81.00	83.72	82.45
HateBERT	83.58	83.00	80.01	84.25	83.93
RoBERTa-Large	84.00	84.00	83.02	85.19	85.87
BERTweet-base	84.00	84.00	82.00	84.73	85.68
DeBERTa-large-v3	86.04	84.5	83.03	86.39	85.47
BERTweet-large	86.55	86.50	83.10	86.07	86.12
Soft Ensemble	86.73	86.01	83.08	86.31	86.00
Hard Ensemble	**86.85**	86.07	83.23	86.19	86.01

Table 1: $F1$-Macro performance for Task A.

Model	w/o DA	with DA
BERT-base	60.33	62.33
BERTweet-base	56.33	59.05
BERT-large	60.66	63.66
RoBERTa	59.33	59.33
RoBERTa-Large	68.00	68.33
DeBERTa-large-v3	68.33	67.16
BERTweet-large	67.33	66.00
Soft Ensemble	69.99	69.00
Hard Ensemble	**70.30**	**70.00**

Table 2: $F1$-Macro performance for Task B.

models already possess more knowledge due to their extensive pre-training. Regarding data augmentation, our findings indicate that doubling all classes resulted in better performance than balancing the dataset.

5.2 Test Phase

We combined the training and development data during the test phase and fine-tuned the models. Our submission, as demonstrated in Table 1, was only made once. We utilized only Module 2 from our pipeline for Task A, employing the hard ensemble strategy. Our three top-performing models, BERTweet-large, DeBERTa-v3-large, and RoBERTa-large, were used without data augmentation. The same approach was adopted for Task B. For Task C, we used RoBERTa-large for the final submission, which yielded the best results in the evaluation set.

5.3 Error Analysis

In this subsection, we have undertaken an error analysis for the submission on Task A. The confusion matrices presented in Figure 4 was constructed to evaluate the performance of our models on the test set. Our ensemble model achieved an $F1$-score of 86.13 on the test set for task A. However, the confusion matrix illustrated in Figure 4 indicates that the model correctly predicted the (not sexist) class 92.80 % of the time (2,813 out of 3,030), while struggling to generate correct predictions for

Model	w/o DA	with DA
BERTweet-base	30.01	30.33
BERT-base	30.00	30.66
RoBERTa-base	34.01	36.66
DeBERTa-large	38.33	38.66
BERT-large	42.33	41.66
BERTweet-large	45.66	45.33
RoBERTa-large	**47.33**	46.66

Table 3: $F1$-Macro performance for Task C.

Table 4: Results on the Test set for All Tasks

Task	Model	F1-score	Rank
A	Ensemble	86.13	10
B	Ensemble	65.50	18
C	RoBERTa-large	46.00	23

the (sexist) class, with a correct prediction rate of only 80.00 % (776 out of 970).

This discrepancy is most likely due to the data imbalance, as 85.7% of the total training set comprises samples labelled as (not sexist). Despite performing data augmentation using back-translation to mitigate the data imbalance issue, the results in the Table1 indicate that this technique did not improve the overall performance. We hypothesise that the back-translation method did not generate diverse samples, and one possible solution is to use data augmentation methods that generate more diverse synthetic data.

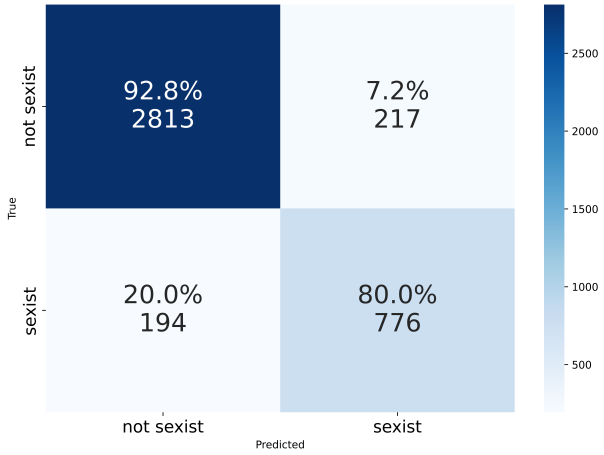

Figure 4: Confusion Matrix for TaskA

6 Conclusion

This paper presents our solution to the Shared Task on Explainable Detection of Online Sexism at SemEval23. Our approach involved employing ensemble voting techniques with previously fine-tuned language models, specifically BERTweet-large, RoBERTa-large, and DeBERTa-V3-large, which resulted in the best performance for both task A and B. Additionally, we discovered that fine-tuning RoBERTa-Large was the most effective approach for addressing task C, outperforming the ensemble voting method. These findings address the first objective of examining how different state-of-the-art transformer-based models perform in sexism detection and classification.

To address our research question; **(RQ): to what extent can data augmentation improve the results and address the data imbalance problem**, we employed a task agnostic data augmentation method, specifically back-translation, in two scenarios: one to double the dataset and the other to augment the underrepresented class. Our results showed that augmenting all classes was more effective than balancing the dataset by augmenting only the underrepresented class, which motivates further exploration of the effects of data augmentation on text classification with unbalanced datasets. In future research, we plan to explore alternative data augmentation techniques to produce more diverse sentences, such as utilizing generative models like GPT-2, to balance and double the dataset's size, and compare the results with the back-translation method.

Moreover, we plan to investigate why augmenting all classes sometimes was more effective than augmenting only the underrepresented class and balancing the dataset.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Tosin Adewumi, Rickard Brännvall, Nosheen Abid, Maryam Pahlavan, Sana Sabah Sabry, Foteini Liwicki, and Marcus Liwicki. 2022. Småprat: Dialogpt for natural language generation of swedish dialogue by transfer learning. In *5th Northern Lights Deep Learning Workshop, Tromsø, Norway*, volume 3. Septentrio Academic Publishing.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Filip Nilsson, Sana Sabah Al-Azzawi, and György Kovács. 2022. Leveraging sentiment data for the detection of homophobic/transphobic content in a multi-task, multi-lingual setting using transformers. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Dymitr Ruta and Bogdan Gabrys. 2005. Classifier selection for majority voting. *Information fusion*, 6(1):63–81.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovács, Foteini Liwicki, and Marcus Liwicki. 2022. Hat5: Hate language identification using text-to-text transfer transformer. *arXiv preprint arXiv:2202.05690*.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. " call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *ICWSM*, pages 573–584.

Rushdi Shams. 2014. Semi-supervised classification for natural language processing. *arXiv preprint arXiv:1409.7612*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jason Wei and Kai Zou. 2019. 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 Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yong Zhang, Hongrui Zhang, Jing Cai, Binbin Yang, et al. 2014. A weighted voting classifier based on differential evolution. In *Abstract and applied analysis*, volume 2014. Hindawi.