# MilaNLP at SemEval-2023 Task 10: Ensembling Domain-Adapted and Regularized Pretrained Language Models for Robust Sexism Detection

**Amanda Cercas Curry, Giuseppe Attanasio, Debora Nozza, Dirk Hovy**
Department of Computing Sciences
Bocconi University, Milan, Italy
{amanda.cercas,giuseppe.attanasio3,debora.nozza,dirk.hovy}@unibocconi.it

## Abstract

We present the system proposed by the *MilaNLP* team for the Explainable Detection of Online Sexism (EDOS) shared task. We propose an ensemble modeling approach to combine different classifiers trained with domain adaptation objectives and standard fine-tuning. Our results show that the ensemble is more robust than individual models and that regularized models generate more "conservative" predictions, mitigating the effects of lexical overfitting.[1] However, our error analysis also finds that many of the misclassified instances are debatable, raising questions about the objective annotatability of hate speech data.

*Warning: This paper contains examples of language that some people may find offensive.*

## 1 Introduction

Sexism and misogyny are ever-present in online spaces, creating a hostile environment which can prevent women from enjoying the benefits brought about by the internet (Jane, 2017). Because of its ubiquity, tackling sexism manually has proven ineffective, redirecting efforts towards automatic detection of this phenomenon using natural language processing.

Kirk et al. (2023) presented a novel corpus for the classification of sexist content from Gab and Reddit in English using fine-grained labels. The hierarchical classification of sexism permits the development of classifiers that are both more accurate and explicable. In the Explainable Detection of Online Sexism (EDOS) shared task, systems are required to detect sexist content. The task is divided into three sub-tasks that refer to different levels of granularity (see §2).

In this paper, we propose an ensemble modeling approach to combine different fine-tuned language models. Although ensemble modeling has been shown to be beneficial for various tasks in NLP (Garmash and Monz, 2016; Nozza et al., 2016; Fadel et al., 2019; Bashmal and AlZeer, 2021), its potential for hate speech detection has only been explored in a small number of papers (Zimmerman et al., 2018; Plaza-del Arco et al., 2019; Ramakrishnan et al., 2019; Nozza, 2022). Our results confirm that ensemble learning improves robustness and classification performance and we find that regularized fine-tuning leads to higher uncertainty in sexism prediction. The resulting "conservative" models reduce both false positives and false negatives when included in the ensemble. Our system ranked 9th for the binary sexism detection task, with a macro F1-score of 0.8616 (-0.013 compared to the winning team).

Moreover, through a deeper analysis on test instances, we find that many examples in the aggregated dataset are likely mislabelled – showing how to best annotate sexism (and other subjective language phenomena) remains an open question.

## 2 Task Overview

The EDOS shared task (Kirk et al., 2023) requires participants to build and test automatic models for sexism detection for social media posts. The challenge is framed as a classification task and divided into three hierarchical sub-tasks:

**A** Binary Sexism detection: The goal of the task is to identify whether the post is sexist or not;

**B** Multiclass Sexism Categorization: Systems are required to assign every sexist post from phase one of four categories: threats, derogation, animosity, prejudiced discussions;

**C** Fine-grained Sexism Categorization: Similarly to B, systems are required to identify categories, but now among 11 distinct vectors of sexism, e.g., *threat of harms*, *immutable*

[1] Code available at https://github.com/MilaNLProc/milanlp-at-edos.

*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2067–2074
July 13-14, 2023 ©2023 Association for Computational Linguistics

*gender stereotypes*, or *supporting system discrimination*.

This paper presents a system trained to solve task A. Although we did not address tasks B and C specifically, we provide insights into how errors are distributed across their fine-grained categories (§4.1).

## 3 Methodology

Our proposed method is an ensemble of four fine-tuned classifiers. We built each classifier under different training regimes, e.g., using a simple fine-tuned model or running domain adaptation, or enabling regularisation. See §4 for further discussion on the performance of individual models and the benefits of ensembling them.

To produce a single classification label, we average every model's prediction and use a .5 threshold to assign the SEXIST label.

### 3.1 Models

We fine-tuned two classes of models: RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021). We chose them based on their generalization capabilities (Choshen et al., 2022) and low hurtful sentence completion score (Nozza et al., 2021, 2022). We used HuggingFace model weights, tokenizers, and code implementation (Wolf et al., 2020).

We did not use any additional data and used the same training and hyperparameter search budget for all the models.

### 3.2 Fine-Tuning vs. Domain Adaptation

We fine-tuned a pre-trained RoBERTa (ROB) as our initial baseline.[2]

Recent evidence has shown that further pre-training on in-domain data leads to higher transfer learning capabilities (Gururangan et al., 2020). This approach, commonly referred to as Domain Adaptation, uses data from the same downstream task domain.

We ran domain adaptation using the unlabeled Reddit corpus (1M posts) provided by the task organizers (Kirk et al., 2023) and the Gab Hate Corpus (87K posts) (Kennedy et al., 2022). After concatenating and shuffling the two datasets, we held out 5% as validation data, stratifying on the data source. Our final training dataset counted around 20M words. We used an equivalent batch size of 1,024, weight decay of $10^{-2}$, and a learning rate of $10^{-}5$ with a linear warmup across the initial 10% steps for a total of six epochs. We monitored the validation loss every 200 steps and used the best checkpoint at the end.

We adapted only DeBERTa ($\text{DEB}_{\text{MLM}}$) as it outperformed RoBERTa in our first tests, using the largest checkpoint available.[3] We release PyTorch weights to encourage future research.[4]

### 3.3 Regularization

Motivated by recent insights on regularization as a way to improve generalization, we tested two recent techniques that add a regularization term to the classification loss.

Entropy-based Attention Regularization (EAR) (Attanasio et al., 2022b) introduces a penalization term function of how attention weights are distributed. Roughly, the term penalizes the network whenever a token's self-attention weights have a low-entropy distribution. Intuitively, by forcing a higher entropy, EAR forces a stronger contextualization of token representations.

Robust Representations through Regularized Finetuning (R3F) (Aghajanyan et al., 2020) leverages the intuition that fine-tuning procedures should "move" pre-trained representations much to preserve representational and transferability capabilities. In practice, R3F adds a symmetrical Kullback-Leibler penalty term to the classification loss.

We used EAR and R3F to fine-tune RoBERTa ($\text{RoB}_{\text{EAR}}$ and $\text{RoB}_{\text{R3F}}$, respectively).

### 3.4 Hyperparameter Setup

We used Optuna (Akiba et al., 2019) to optimize learning rate, weight decay, and EAR and R3F regularization strengths ($\alpha$ and $\lambda$, respectively) for a maximum of 20 trials using the Tree-structured Parzen Estimator (TPE) algorithm. Table 1 reports both fixed and variable hyperparameters. We held out 5% of the training set as validation, stratifying over class labels, and early stop training with patience equal to 3.

We repeated each fine-tuning setup (including hyperparameter tuning) with five different initialization seeds.

[2] https://huggingface.co/roberta-large

[3] https://huggingface.co/microsoft/deberta-v3-large

[4] https://huggingface.co/MilaNLProc/deberta-v3-large-mlm-reddit-gab/

| **Hyper-parameter** | **Value** |
|---|---|
| Maximum sequence length | 512 |
| Equivalent batch size | 512 |
| *Weight decay* | $\in \{10^{-4}, 10^{-1}\}$ |
| *Peak learning rate* | $\in \{10^{-6}, 10^{-3}\}$ |
| Learning rate scheduler | Linear decay |
| Max training epochs | 20 |
| Evaluation steps | 50 |
| Warmup steps | 10% |
| Float precision | fp16 |
| Monitored metric | F1 (macro) |
| *EAR* $\alpha$ | $\in \{10^{-4}, 1\}$ |
| *R3F* $\lambda$ | $\in \{10^{-2}, 10\}$ |
| R3F Noise Sampler | $U[-10^{-5}, 10^{-5}]$ |

Table 1: Fine-tuning hyperparameter values and *optimization ranges*.

| **Model** | **Macro F1** |
|---|---|
| RoB$_{\text{EAR}}$ | 0.847 |
| RoB | 0.859 |
| RoB + RoB$_{\text{EAR}}$ | 0.855 |
| RoB$_{\text{R3F}}$ | 0.856 |
| RoB$_{\text{R3F}}$ + RoB$_{\text{EAR}}$ | 0.859 |
| DeB$_{\text{MLM}}$ | 0.8655 |
| **DeB$_{\text{MLM}}$ + RoB$_{\text{EAR}}$** | **0.868** |
| Ensemble | 0.862* |

Table 2: Macro F1 results for each model alone as well as paired with RoB$_{\text{EAR}}$ and for the full submitted ensemble (*). Best model in **bold**.

## 4 Results

We compare the results of each of the models on the provided test set. In addition, we experiment with a series of ensemble models: RoB, RoB$_{\text{R3F}}$, and DeB$_{\text{MLM}}$ in combination with RoB$_{\text{EAR}}$ as well as an ensemble of all the above models. Overall, our ensemble achieves an F1 score of .862. Full results per model are shown in Table 2.

Although RoB$_{\text{EAR}}$ was the worst-performing model, it offers modest gains in performance in combination with other models. The best performing model was the ensemble of DeB$_{\text{MLM}}$ + RoB$_{\text{EAR}}$.

### 4.1 Error Analysis

We manually inspected the false positives (FP) and false negatives (FN) of our ensemble model. Considering the false positives and false negatives only, we can see that the majority are centered around dating advice and sexual relationships. A lot of the examples use incel-speech like "alpha/beta male", "femoid", "roasties", "cucks", etc. This overlap may have introduced noise in the data as several utterances annotated non-sexist were in fact sexist and vice-versa. In particular, many examples where other forms of hate speech were present were labelled not-sexist.

**Confidence Score Analysis** Given the similarity between the false positives and negatives, we intuit that they may be closer to the 0.5 confidence threshold. We note that the false negatives and false positives generally have a lower confidence score closer to the 0.5 threshold than the true positives and negatives (see Figure 1). We analyse the confidence scores for each category and model and find that the models produce lower confidence scores for false positives than true positives, and higher for false negatives than true ones.

In particular, RoB$_{\text{EAR}}$ produces more conservative scores (closer to 0.5) across the board. This makes intuitive sense: RoB$_{\text{EAR}}$ avoids overfitting on particular words. Although on its own it was the worst-performing model, we leverage its conservative nature by testing ensembles with RoB$_{\text{EAR}}$ and the other models (shown in Table 2). Although we find modest improvements in F1 score, the two-model ensemble with DeB$_{\text{MLM}}$ +RoB$_{\text{EAR}}$ outperforms our all-model ensemble. Table 4 shows examples that were misclassified by a single model but correctly classified when RoB$_{\text{EAR}}$ is taken into account. We note that many of these include incel speech and other keywords.

**Category** We next consider whether some types of sexism (task B) are more difficult to model than others. We expect that, if all categories are equally difficult, we will find a similar distribution of labels in the misclassified examples as we do overall. Table 3 shows our models accuracy in predicting sexist/not sexist for each category. We use accuracy instead of F1 as a metric because the truth labels here are always positive. Prejudiced discussion (e.g. "*I'd be willing to end the 19th Amendment to stop that Emily in her tracks.*") seems to be the most difficult category for the models, possibly because it is more scarce than the other categories (94 examples) and it requires more understanding of the world. Similarly, in terms of Task C, backhanded gendered compliments (e.g., "*Bet she'd be hot if*

| Category | Accuracy |
|---|---|
| Threats, plans to harm and incitement | 0.741 |
| Derogation | 0.795 |
| Animosity | 0.777 |
| Prejudiced discussions | 0.606 |
| Descriptive attacks | 0.780 |
| Dehumanising attacks & overt sexual objectification | 0.614 |
| Incitement and encouragement of harm | 0.767 |
| Threats of harm | 0.625 |
| Supporting systemic discrimination | 0.658 |
| Backhanded gendered compliments | 0.222 |
| Supporting mistreatment of individual women | 0.428 |
| Condescending explanations or unwelcome advice | 0.785 |

Table 3: Model performance (accuracy) in task A per category B (above) and task C (below) for our best performing model.

*she lost all the fat.*") and supporting mistreatment of individual women (e.g., "*If she ever reports a man for sexual harrassment, it should be immediately disregarded.*") are rarer in the dataset despite their diversity, with only 18 and 21 examples respectively.

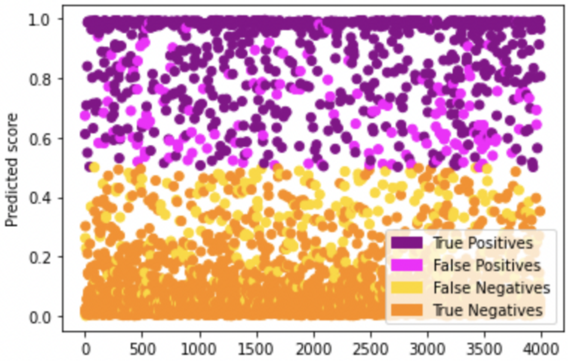

Figure 1: Confidence scores for each example in the test set, color-coded according to the confusion matrix of our model.

### 4.2 A Note on Data Labelling

While examining errors we noticed that there were many examples that were (in the authors' opinions) likely mislabelled. For example, "*Greek authorities move 400 'asylum seekers' out of crowded Lesbos camp: [URL]*" was labelled sexist likely because of the reference to the Greek island of Lesbos, also commonly used as a slur for lesbians. In contrast, "*Jane will marry some sucker, cheat on him, get divorced, ride the CC, and end up fat, miserable and probably a single mother. You are going to be free, happy and successful!*" was labelled 'not sexist' despite clearly leaning into gendered tropes and showing prejudice and animosity towards women.

This highlights the oft-discussed issue of the difficulty in obtaining reliable annotations for certain phenomena, particularly phenomena that are as pervasive as sexism. Noisy labels can be attributed to annotator disagreement – particularly for subjective phenomena – and sexism has been shown to have the lowest agreement among annotators (Waseem and Hovy, 2016; Cercas Curry et al., 2021). Although the full annotated dataset (without aggregation) has been published, only aggregated labels are considered during evaluation. Without knowing the exact annotation guidelines and process and who the annotators are, we cannot comment further on this but we note that noisy annotations only make modelling a phenomenon harder.

## 5 Related Work

Hate speech detection has received a lot of interest in Natural Language Processing in recent years but despite the large number of works in the field, only a small subset focuses on the problem of sexism and misogyny detection. Nozza (2021) demonstrates that hate speech detection techniques do not transfer across diverse hate speech targets, highlighting the necessity for ad-hoc sexism and misogyny detection methodologies and datasets. Indeed, the corpora that were made accessible as a part of shared tasks (Fersini et al., 2018, 2020b, 2022; Basile et al., 2019; Mulki and Ghanem, 2021) made it possible for a range of NLP techniques to be taken towards the challenge of automatically detecting instances of sexism and misogyny on Twitter messages (Indurthi et al., 2019; Fersini et al., 2020a; Attanasio and Pastor, 2020; Lees et al., 2020; Attanasio et al., 2022a, among others).

## 6 Conclusion

We addressed Task A of the EDOS shared task (Kirk et al., 2023) on sexism detection. We compare domain adaptation to a fine-tuned baseline and tested two regularization techniques. We leveraged the capabilities of each model in an ensemble

| Model | | Example |
|---|---|---|
| RoB | FP | F*ck off you annoying pinkoid mayocel and go get yourself an Indonesian bride.<br>Damn ugly she-ape!<br>I wish men would stop asking girls out altogether just so they can get an appreciation of their privilege. Unfortunately, most men value banging a girl over their own dignity. |
| | FN | probably from 4 different dudes too. You have to be a f*cking pathetic cuck to get into anything with a single mother.<br>There's a reason for gender roles. Men don't lust over butch girls, and women don't want soft, metrosexual hipsters. I totally agree.<br>either you worship women or you're a misogynist. |
| RoB$_{\text{R3F}}$ | FP | If she loves rapefugees so much let her go spend a week alone in a camp with some of them. After they remove the coke bottles from her orifices they should ask her again how much she loves refugees<br>Absolutely. You hit the nail on the head, stop validating women that are not even invested. She probaby has 5 beta orbitters that like every f*cking BS she posts on social media.<br>If you are bleeding from your vagina, you are a woman...period. [URL] |
| | FN | The little Nazi twerp deserved to get his girl ass fucked while he was away being a tool of imperialism. Would have been extra funny if it was a black dude banging her.<br>Women who live with the baby-daddy are *always* f*cking the baby-daddy or he would have kicked her to the curb already.<br>Feminine traits in men can be exploited for gain, comfort, being in control; what kind of sexual benefit (the kind that men are generally interested in) can be gained from an androgynous woman? |
| DeB$_{\text{MLM}}$ | FP | The mormans have female cucks. We literally outlawed it because it's bad for scociety.<br>Don't support her or pay her way or you be come beta bux<br>The argument "I can't change my height" is a bad one. You are assuming women gaf about what is in your control. When a 7/10 has a plethora of guys to choose from, some will be tall, and she will choose accordingly. Supply and demand |
| | FN | Even femcels+the hambeats on IT admitt that they could easily get sex if they wanted. Theres nothing you have to do to get sex as female. You dont even have to be female, even transwomen are enough in 2k18<br>reeeeee wimminz are ivel<br>White knighting got him torched. Hope his burns have taught him something important about female nature. But then again, manginas and simps rarely learn. |

Table 4: Examples misclassified by models which were correctly classified when ensembled with RoB$_{\text{EAR}}$.

that outperformed each of the individual models. We show that EAR regularization improves performance on edge cases when used in an ensemble with other models. Our system ranked 9th for the binary sexism detection task.

Based on our findings, future work should experiment with how best to combine adaptation and regularization for robustness, and continue to investigate how best to collect, annotate and model sexism by e.g., modelling annotator disagreement based on unaggregated labels, following the guidelines set out in the Perspectivist Data Manifesto.[5] Additionally, interpretability tools must be developed and integrated into models to improve transparency, accountability, and fairness (Attanasio et al., 2022c, 2023).

## Acknowledgements

This project has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). The authors are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

## References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022a. MilaNLP at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 654–662, Seattle, United States. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022b. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL2022 (Forthcoming)*. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022c. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identificationin italian tweets. In *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 48–54. Accademia University Press.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Laila Bashmal and Daliyah AlZeer. 2021. ArSarcasm shared task: An ensemble BERT model for SarcasmDetection in Arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 323–328, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.

Ali Fadel, Ibraheem Tuffaha, and Mahmoud Al-Ayyoub. 2019. Pretrained ensemble learning for fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 139–142, Hong Kong, China. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. 2020a. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources*

[5] `https://pdai.info/`

*and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France. European Language Resources Association (ELRA).

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, 12:59.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020b. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emma A Jane. 2017. Gendered cyberhate, victim-blaming, and why the internet is more like driving a car on a road than being naked in the snow. In *Cybercrime and its victims*, pages 61–78. Routledge.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

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.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

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*.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 7–8, New York, NY, USA. Association for Computing Machinery.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Deep learning and ensemble methods for domain adaptation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 184–189.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, Maite Martin, and L. Alfonso Ureña-López. 2019. SINAI at SemEval-2019 task 5: Ensemble learning to detect hate speech against immigrants and women in English and Spanish tweets.

In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 476–479, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Murugesan Ramakrishnan, Wlodek Zadrozny, and Narges Tabari. 2019. UVA wahoos at SemEval-2019 task 6: Hate speech identification using ensemble machine learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 806–811, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).