UL & UM6P at SemEval-2023 Task 10: Semi-Supervised Multi-task Learning for Explainable Detection of Online Sexism

Salima Lamsiyah[1], **Abdelkader El Mahdaouy**[2], **Hamza Alami**[3]
Ismail Berrada[3] and **Christoph Schommer**[1]
[1]Dept. of Computer Science, Faculty of Science, Technology and Medicine,
University of Luxembourg, Luxembourg
[2]Modeling, Simulation and Data Analysis, Mohammed VI Polytechnic University, Morocco
[3]School of Computer Science, Mohammed VI Polytechnic University, Morocco
{firstname.lastname}@{uni.lu[1], um6p.ma[2,3]}

Abstract

This paper introduces our participating system to the Explainable Detection of Online Sexism (EDOS) SemEval-2023 - Task 10: Explainable Detection of Online Sexism. The EDOS shared task covers three hierarchical sub-tasks for sexism detection, coarse-grained and fine-grained categorization. We have investigated both single-task and multi-task learning based on RoBERTa transformer-based language models. For improving the results, we have performed further pre-training of RoBERTa on the provided unlabeled data. Besides, we have employed a small sample of the unlabeled data for semi-supervised learning using the minimum class-confusion loss. Our system has achieved macro F1 scores of 82.25%, 67.35%, and 49.8% on Tasks A, B, and C, respectively.

1 Introduction

The advent of diverse social media platforms has enabled both identified and anonymous users to express their views and beliefs freely and openly. However, a number of online users leverage this freedom of expression to abuse other people or a group of people based on their race, ethnicity, culture, religion, gender, and political orientation, to name but a few (Guiora and Park, 2017). To tackle this issue, social media platforms rely on content moderation techniques to review and filter their users' posts (Shen and Rose, 2019; Gillespie, 2018, 2020; Liu et al., 2021). During the past decade, researchers have shown an increased interest in building artificial intelligence-based tools and applications to address the challenges of content moderation automation in social media platforms. For instance, various Natural Language Processing (NLP) research works have been introduced for offensive language (Zampieri et al., 2019, 2020), hate speech (Basile et al., 2019; Röttger et al., 2021), Cyberbullying (Van Hee et al., 2015; Menini et al., 2019; Chen and Li, 2020), toxicity (van Aken et al., 2018; Pavlopoulos et al., 2021), misogyny (Fersini et al., 2018; Mulki and Ghanem, 2021) and sexism (Jha and Mamidi, 2017; Rodríguez-Sánchez et al., 2021) detection.

The Explainable Detection of Online Sexism (EDOS) shared task introduces three sub-tasks for sexism detection and categorization in the English language (Kirk et al., 2023). The aim is to develop NLP models and systems for sexism detection and explainability. It emphasizes improving predictions' interpretability and explainability for fair content moderation decision-making. The EDOS shared task provides training data for developing an accurate and explainable model for sexism detection using a set of hierarchical labels and tasks.

In this paper, we present our participating system to the EDOS shared task. We have explored both single-task and multi-task learning models. All our models rely on RoBERTa pre-trained transformer-based language model (Liu et al., 2019). To improve the performance of our models, we have conducted domain-adaptive pre-training using the shared task unlabeled data. To do so, on the one hand, we have employed the whole word masking pre-training objective (Cui et al., 2019). On the other hand, we have performed semi-supervised training using the provided labeled data and a sample of 28k entries from the unlabeled data, predicted as sexist texts by our ST_3 model (Section 3.3), on Task B and Task C. The Minimum Class Confusion (MCC) loss (Jin et al., 2019) is then used to train our model on the sampled unlabeled data.

For the official evaluation, we have submitted the results of our multi-task learning model which is trained in a semi-supervised manner and utilizes our adapted RoBERTa-large encoder. Our system has achieved macro F1 scores of 82.25%, 67.35%, and 49.8% on Tasks A, B, and C, respectively.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 644–650
July 13-14, 2023 ©2023 Association for Computational Linguistics

Table 1: Labels hierarchy of EDOS sub-tasks

Task A	Task B	Task C
Not Sexist	−−	−−
Sexist	Threats, plans to harm and incitement	Threats of harm
		Incitement and encouragement of harm
	Derogation	Descriptive attacks
		Aggressive and emotive attacks
		Dehumanising attacks and overt sexual objectification
	Animosity	Causal use of gendered slurs, profanities and insults
		Immutable gender differences and gender stereotypes
		Backhanded gendered compliments
		Condescending explanations or unwelcome advice
	Prejudiced Discussions	Supporting mistreatment of individual women
		Supporting systemic discrimination against women as a group

2 Background

2.1 Task Description

The Explainable Detection of Online Sexism (EDOS) shared task covers three sub-tasks for sexism detection and categorization in the English language (Kirk et al., 2023). The aim is to develop models that detect sexist content and explain why it is sexist. It addresses the challenges of developing an accurate and explainable model for sexism detection using a set of hierarchical labels and tasks. Table 1 illustrates the class-label hierarchy of EDOS sub-tasks. These sub-tasks are described as follows:

- **Task A - Binary Sexism Detection** is a binary classification that aims to detect sexist posts.
- **Task B - Category of Sexism** is a multi-class classification task. It aims to assign a sexist post to one of the following categories: (1) threats, (2) derogation, (3) animosity, and (4) prejudiced discussions.
- **Task C - Fine-grained Vector of Sexism** is a multi-class classification task. The goal is to assign a sexist post to more fine-grained sexism vector categories (see table 1) .

The shared task data is collected from Gab and Reddit. The task-labeled dataset consists of 20K data instances (entries), where half of the entries are sampled from Gab and the other half from Reddit. The original dataset is split following the 70/10/20 proportions for training/validation/test sets. In addition to the labeled dataset that is provided for models training on the downstream tasks, the organizers have supplied two unlabelled datasets containing 2M text entries (1M entries from GAB and 1M from Reddit).

2.2 Related Work

Online sexism is a pervasive problem that can harm women and create hostile environments. Sexism detection in social media content has become an emerging field of natural language processing and social computing (Jha and Mamidi, 2017; Karlekar and Bansal, 2018; Zhang and Luo, 2019; Parikh et al., 2019; Abburi et al., 2020; Chiril et al., 2020; Rodríguez-Sánchez et al., 2021; Sen et al., 2022). In the *Automatic Misogyny Identification* (AMI) shared task at IberEval and EvalIta 2018, participants were tasked with identifying instances of sexist behavior in tweets and categorizing them based on a taxonomy proposed by Anzovino et al. (2018). The latter includes behaviors such as discredit, stereotype, objectification, sexual harassment, threat of violence, and derailing. Most participating systems have employed the SVM and ensemble classifiers with features such as n-grams and opinions (Fersini et al., 2018). The AMI shared task's datasets have also been utilized in the *Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* shared task at SemEval 2019, where the best results have also been achieved by an SVM model utilizing sentence embeddings as features (Indurthi et al., 2019).

Deep learning-based methods have also shown encouraging performance for sexism detection. For instance, Jha and Mamidi (2017) have used an LSTM model for detecting and classifying tweets as benevolent, hostile, or non-sexist. Karlekar and Bansal (2018) have used a single-label CNN-LSTM model with character-level embeddings to classify three types of sexual harassment, namely commenting, ogling/staring, and touching/groping. Zhang and Luo (2019) have employed two different types of deep neural network models, namely CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction, intending to categorize social media texts into one of three categories: racist, sexist, or non-hateful. Additionally, Parikh et al. (2019) have analyzed instances of sexism reported by women on the "*Everyday Sexism Project*" website and categorized them into 23 non-mutually exclusive categories using a variety of deep learning models, including LSTM, CNN, CNN-LSTM, and BERT. These models were trained on top of different distributional representations such as characters, subwords, words, and sentences, as well as additional linguistic features. In the same context, Chiril et al. (2020) have introduced a French sexism detection method based on BERT contextualized word embeddings complemented with both linguistic features and generalization strategies. Abburi et al. (2020) have proposed a semi-supervised multi-task learning method using BERT model for multi-label fine-grained sexism classification. Another line of work uses counterfactually augmented data to improve out-of-domain generalizability for sexism and hate speech detection (Sen et al., 2022). Finally, Rodríguez-Sánchez et al. (2021) have organized the *sEXism Identification in Social neTworks* (EXIST) shared task at IberLEF, which involves sexism identification and categorization of tweets and gabs in both Spanish and English.

In addition to detecting online sexism, several research works have been proposed to address associated societal issues such as hate speech (Basile et al., 2019; Röttger et al., 2021), offensive language (Zampieri et al., 2019, 2020), cyberbullying (Van Hee et al., 2015; Menini et al., 2019; Chen and Li, 2020), misogyny (Fersini et al., 2018; Mulki and Ghanem, 2021), and toxicity (van Aken et al., 2018; Pavlopoulos et al., 2021). However, most existing approaches are based on opaque deep learning models whose inner workings cannot easily be explained. These models only provide binary or coarse-grained labels that fail to provide insights into the specific types of sexism present in a given text or the reasoning behind its classification. Therefore, in recent years, there has been a growing interest in the field of NLP to develop models that not only make accurate predictions but also provide explanations of how they reached their decisions (Danilevsky et al., 2020; Balkir et al., 2022; Kim et al., 2022).

3 System Overview

In this section, we present the employed text encoders, the pre-training procedure, and our model architectures.

3.1 Text Encoder

To encode the input texts of EDOS sub-tasks, we have explored both RoBERTa *base* and *large* variants. RoBERTa is a Pre-trained Language Model (PLM) based on the transformer encoder architecture (Liu et al., 2019). It is a variant of BERT model (Devlin et al., 2019) trained using an optimized approach. More precisely, it is pre-trained on five English text corpora of varying sizes and domains: Book Corpus, CC-News, OpenWeb Text, and Stories datasets (Liu et al., 2019). The authors improve the pre-training procedure of BERT by increasing the number of training steps, using bigger batch sizes, and employing larger pre-training data. Besides, they have omitted the next sentence prediction objective and used longer training sequences (Liu et al., 2019).

3.2 Further pre-training

To adapt the language model to Gab and Reddit domain data, we have conducted domain adaptive fine-tuning of RoBERTa PLM using the provided starter-kit unlabeled data of EDOS shared task. Further pre-training is performed by duplicating the unlabeled data (three times) and optimizing the whole word masking pre-training objective (Cui et al., 2019).

3.3 Models

We have assessed the performance of three single-task and three multi-task learning models. The single-task models are described as follows:

- **ST_1**: This model consists of RoBERTa-base encoder and one dropout and one classification layer.

- **ST_2**: This model has a similar architecture to ST_1 model, while using RoBERTa-large to encode the input texts.

- **ST_3**: This model is also similar to the previous ones, but it relies on our adapted RoBERTa-large PLM to encode the input texts.

In our three multi-task learning models, we have utilized our adapted RoBERTa-large language model (further pre-training of RoBERTa-large on Gab and Reddit unlabeled data) to encode the input texts. These models are described as follows:

- **MT_1**: This model uses a classifier per task on the embedding of our adapted RoBERTa-large PLM. Each classifier consists of one dropout layer and one classification layer.

- **MT_2**: This model applies task-specific attention layers on the contextualized embedding of our RoBERTa large. The aim is to extract task-discriminative features. This model uses also similar three classifiers that consist of one dropout and one classification layer. Nevertheless, the input of each classifier is the concatenation of the task-specific attention output and the pooled output embedding. This model architecture has been used in several previous shared tasks, including the Arabic misogyny detection and categorization (El Mahdaouy et al., 2021; Essefar et al., 2021; Mahdaouy et al., 2021).

- **MT_3**: This model is similar to MT_2, but it is trained in a semi-supervised manner using the Minimum Class Confusion (MCC) loss (Jin et al., 2019) on tasks B and C, respectively. To do so, we have employed a sample of 28k instances (14k from Gab and 14k from Reddit data) of the starter-kit unlabeled data that are predicted sexist by the ST_3 model. The Cross-Entropy (CE) losses (tasks A, B, and C) are minimized using the provided labeled data and the MCC losses (tasks B and C) are optimized on the sampled unlabeled data. Thus, the total loss combines 5 training objectives. In order to weight the five losses in the overall objective, we have used the automatically weighted multi-task loss (Kendall et al., 2018).

4 Experimental Setup

All our models are implemented using Pytorch[1] deep learning framework, Pytorch Lightning[2], and Hugging Face Transformers[3] library. We have conducted our experiments using Dell PowerEdge C4140 server, having 4 Nvidia V100 SXM2 32GB.

We have performed domain adaptive pre-training using a learning rate of 5×10^{-5} and a batch size of 8 per GPU device. The number of epochs is fixed to 3. The other hyper-parameters are fixed to their defaults values of the employed pre-training script[4].

For model fine-tuning on the EDOS sub-tasks, we have fixed the learning to 1×10^{-5}, the dropout to 0.2, the maximum sequence length to 128, and the number of epochs to 10. The batch size is fixed to 16 for all sub-tasks. All models have been trained on the training set and validated on the provided development set. We have performed model evaluation using the macro averaged Recall, Precision, and F1-score.

5 Results

In this section, we describe the obtained results of our models as well as our official submissions. Table 2 summarizes the obtained results of our models on the development sets of EDOS sub-tasks. For all sub-tasks, we present the obtained results on the macro-averaged Precision, Recall, and F1-score.

Task A

On the one hand, for single-task learning models, the obtained results on Task A show that RoBERTa-large model (ST_2) outperforms its base variant (ST_1). Besides, domain-adaptive pre-training (ST_3) improves the performance of our model. On the other hand, the obtained results using multi-task learning models demonstrate that using task-specific attention layers (MT_2) enhances the performance of our system. Although the MCC training objective is minimized on Tasks B and C, the best F1-score on Task A is obtained using the MT_3 model.

[1] https://pytorch.org/
[2] https://www.pytorchlightning.ai/
[3] https://github.com/huggingface/transformers
[4] https://github.com/huggingface/transformers/blob/main/examples/research_projects/mlm_wwm/run_mlm_wwm.py

Table 2: The obtained results (%) on the development sets of EDOS subtasks.

	Task A			Task B			Task C		
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ST_1	82.94	80.73	81.73	61.49	59.82	59.82	45.71	40.19	41.43
ST_2	83.50	83.45	83.48	65.13	64.86	64.96	49.13	48.02	48.26
ST_3	**85.86**	82.68	84.09	68.99	69.38	69.18	51.54	50.35	50.26
MT_1	80.67	84.92	82.34	**69.22**	70.88	69.98	52.05	52.64	51.41
MT_2	81.27	84.97	82.79	68.44	73.35	70.47	51.29	55.28	52.63
MT_3	83.34	**86.62**	**84.74**	68.31	**74.8**	**70.81**	**54.17**	**57.79**	**55.19**

Task B

In line with the obtained results on Task A, employing RoBERTa-large and domain-adaptive pretraining (ST_2) improves the categorization performance of our single-task learning model. Besides, the comparison results demonstrate that our multi-task learning models surpass their single-task learning counterparts on most evaluation measures. The best F1-score performance is obtained using MT_3 model that minimizes the MCC loss on Task B, and Task C.

Task C

In accordance with Task A and Task B, domain-adaptive pretraining enhances the performance of our single-task learning model (ST_3). Overall, the multi-task learning models improve the F1-score results. Besides, using task-specific attention layers improves the model performance. Finally, the best results are obtained using MT_3 model that minimizes the MCC loss on Task B, and Task C.

Official submissions results

For the official evaluation results, we have submitted the results of our multi-task learning model MT_3. Table 3 present our official results on EDOS sub-tasks. The official evaluation results show that our system achieved macro F1 scores of 82.25%, 67.35%, and 49.8% on Tasks A, B, and C, respectively.

Table 3: The official results (%) of our submitted system.

	Task A	**Task B**	**Task C**
F1-score	82.25	67.35	49.80

6 Conclusion

In this paper, we have introduced our participating system to the Explainable Detection of Online Sexism (EDOS) shared task. In order to improve the performance of our model, we have explored domain-adaptive pre-training and semi-supervised learning leveraging the provided unlabeled data. Overall, we have assessed the performance of three single-task learning models and three multi-task learning models.

The overall evaluation results show that our system has achieved promising results on Task B and Task C. It is ranked 11th, and 9th on Task B, and Task C, respectively.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2020. Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5810–5820, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Esma Balkir, Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2672–2686, Seattle, United States. 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.

Hsin-Yu Chen and Cheng-Te Li. 2020. HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2543–2552, Online. Association for Computational Linguistics.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese BERT. *CoRR*, abs/1906.08101.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

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.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 334–339, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kabil Essefar, Abdellah El Mekki, Abdelkader El Mahdaouy, Nabil El Mamoun, and Ismail Berrada. 2021. CS-UM6P at SemEval-2021 task 7: Deep multi-task learning model for detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1135–1140, Online. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples*. Accademia University Press.

T. Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Amos Guiora and Elizabeth A Park. 2017. Hate speech on social media. *Philosophia*, 45:957–971.

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.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. 2019. Minimum class confusion for versatile domain adaptation.

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811, Brussels, Belgium. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Yi Liu, Pinar Yildirim, and Z. John Zhang. 2021. Social media, content moderation, and technology.

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*.

Abdelkader El Mahdaouy, Abdellah El Mekki, Ahmed Oumar, Hajar Mousannif, and Ismail Berrada. 2021. Deep multi-task models for misogyny identification and categorization on arabic social media. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 852–860. CEUR-WS.org.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy. Association for Computational Linguistics.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). In *Forum for Information Retrieval Evaluation*, pages 7–8.

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. 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 1642–1652, Hong Kong, China. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

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.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.