Coco at SemEval-2023 Task 10: Explainable Detection of Online Sexism

Kangshuai Guo♠, Ruipeng Ma ♠, Shichao Luo♣, Yan Wang♠ *
♣China CITIC Baixin Bank Corporation Limited
♠University of Electronic Science and Technology of China
{202122080434,202221080135}@std.uestc.edu.cn
xiaoluoyfy@gmail.com,yanbo1990@uestc.edu.cn

Abstract

Sexism is a growing concern on social media platforms because it affects the overall health of the internet and can have negative impacts on society. This paper describes the coco system that participated in SemEval-2023 Task 10, Explainable Detection of Online Sexism (EDOS), which aims at sexism detection in various settings of natural language understanding. We develop a novel neural framework for sexism detection and misogyny that can combine text representations obtained using pre-trained language model models such as Bidirectional Encoder Representations from Transformers and using BiLSTM architecture to obtain the local and global semantic information. Further, considering that the EDOS dataset is relatively small and extremely unbalanced, we conducted data augmentation and introduced two datasets in the field of sexism detection. Moreover, we introduced Focal Loss which is a loss function in order to improve the performance of processing imbalanced data classification. Our system achieved an F1 score of 78.95% on Task A - binary sexism.

1 Introduction

Social media platforms offer unparalleled communication and information-sharing abilities while providing an anonymous space for people of varying genders, ethnicities, races, and cultures to interact online. (Fortuna et al., 2021; Pamungkas et al., 2021; Chiril et al., 2020; Williams et al., 2020). As a result, there has been a rise in incidents, hostile behavior, and instances of harassment on social media platforms, especially sexism (Kirk et al., 2023; Deluca et al., 2016). Sexism is connected to beliefs about the essential nature of women and men, as well as the roles they should assume in society. This hierarchical way of thinking can be deliberate and hostile, or it can be unintentional and take the form of unconscious bias.

The SemEval task 10 supports the development of English-language models for sexism detection that fine-grained classifications for sexist content from Gab and Reddit, which consists of three subtasks (Kirk et al., 2023). As illustrated in Fig.1:

(1) Task A - Binary Sexism Detection: It is a binary classification task which models are required to predict whether a text contains gender bias or not;

(2) Task B - Category of Sexism: It is a four-class classification task on texts that exhibit gender bias which requires models to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions;

(3) Task C - Fine-grained Vector of Sexism : It is an 11-class classification task for texts exhibiting gender bias which requires models to predict one of 11 fine-grained vectors.

Our method employed Bert (Devlin et al., 2018) as mono-lingual pre-trained language models, and various models including CNN, LSTM and BiLSTM for further feature extraction to obtain context semantic features. Our contributions are summarized below.

- Our system achieved the Macro F1 score of 78.95% on the Task A - binary sexism detection;
- we employ a suitable neural framework for sexism detection task that combines using a pre-trained language model Bert and BiLSTM architecture.
- we introduced the additional datasets including EXIST-2022 and TRAC-2020 for data augmentation and tried different loss functions in order to solve the label imbalance problem, to achieve better performance for the sexism detection task.

*: Corresponding author.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 469–476

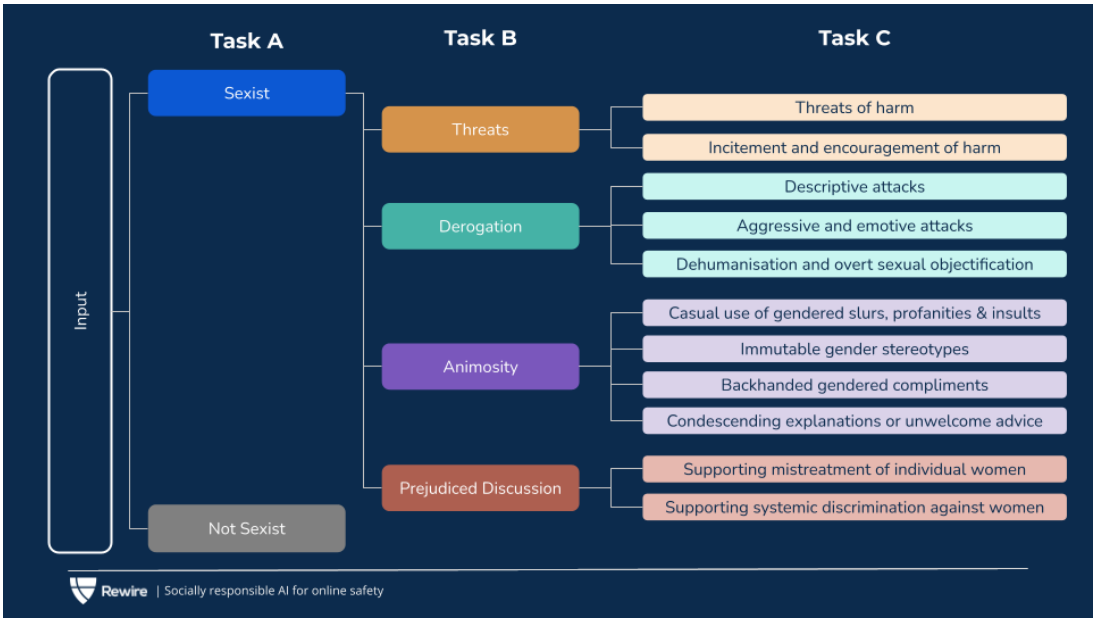

Figure 1: Description of subtasks about Explainable Detection of Online Sexism.

2 Background

In this section, we review the work on the classification of sexism or misogyny. We also pay attention to some work related to the identification of hate speech, because some of these works are applicable to sexism detection to some extent, and these methods treat sexism as a kind of hate (Davidson et al., 2017; Abburi et al., 2021).

Over the last few years, due to the exponential growth of user-generated content and the diverse range of behaviors towards women on social media, manually inspecting and moderating sexist content has become unfeasible (Jiang et al., 2022). Therefore, there has been a significant surge in academic research aimed at automatically detecting sexist behavior in both monolingual and multilingual scenarios in recent years (Jha and Mamidi, 2017; Rodríguez-Sánchez et al., 2020). To tackle this issue, the first attempt was made by Hewitt et al.(Hewitt et al., 2016), who studied the manual detection of sexist tweets. Anzovino et al. (Anzovino et al., 2018) conducted the first survey of automatic sexism detection in social media. In an effort to address the issue of unintended bias in machine learning models for sexism detection, a new method proposed by Nozza et al. (Nozza et al., 2019) try to measure and mitigate this bias. Agrawal et al. (Agrawal and Awekar, 2018) investigated the detection of cyberbullying on social media platforms using deep learning methods. Unlike previous studies on hate detection, they attempt to identifying sexism and misogyny, which encompass hate speech targeted towards women but is not limited to hate (Abburi et al., 2021).

Recently, academic work research on the detection of sexism or misogyny has been conducted extensively, particularly in multilingual and cross-domain scenarios (Pamungkas et al., 2020; Parikh et al., 2021). The majority of studies focusing on the detection of sexism or misogyny against women utilize supervised approaches. In their research, Basile et al. (Basile and Rubagotti, 2018) used the support vector machine to perform sexism detection, with features based on word and character n-grams. Buscaldi et al.(Buscaldi, 2018) employed random decision forests which is an ensemble learning method by constructing a multitude of decision trees at training and utilized locally weighted character n-grams as features for the detection of sexism or misogyny. Deep learning approaches based on Pre-Trained Models(PTM) have become more popular, such as DeBERTa (He et al., 2020), Roberta (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2019), XLNet (Yang et al., 2019), and SpanBERT (Joshi et al., 2020), which have made state-of-the-art achievements in different languages (Chiril et al., 2020; Samory et al., 2021). Parikh et al.(Parikh et al., 2019) built a BERT-based neural architecture to perform sexism detection and showed the models perform better than classic supervised approaches.

There are many opensource multilingual datasets

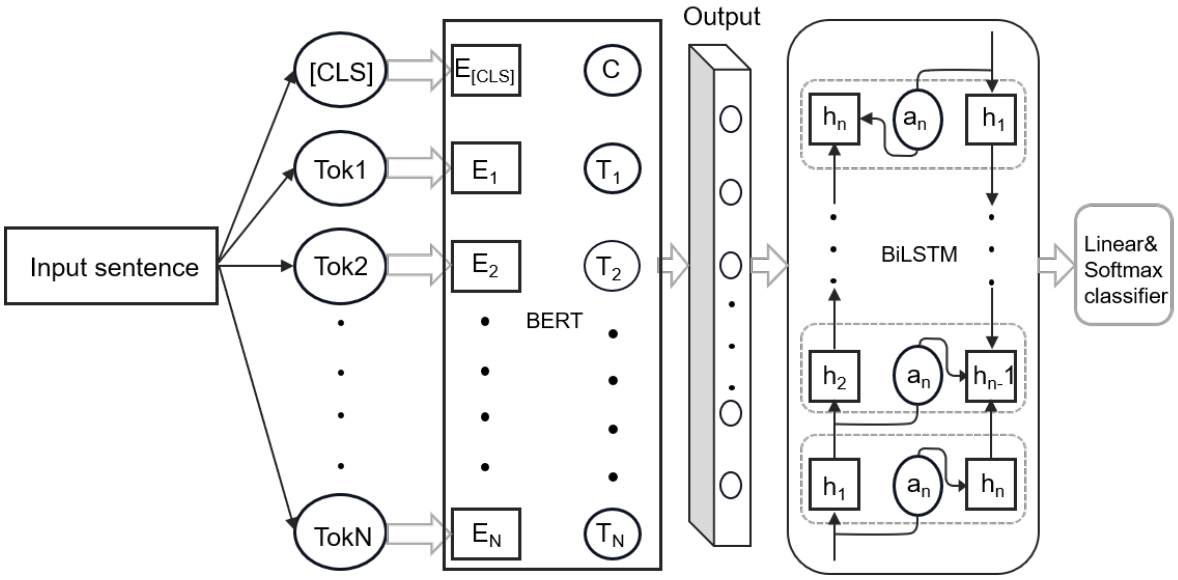

Figure 2: The structure of BERT-BiLSTM. The model we propose comprises the following components: Bert, BiLSTM, and The output layer, which comprises a fully connected layer and a softmax layer.

in the field of detection of sexism or hate speech, such as EXIST-2022 (Rodríguez-Sánchez et al., 2022), TRAC-2020 (Bhattacharya et al., 2020), HatEval (Basile et al., 2019) and AMI@IberEval (Fersini et al., 2018). Our model introduced the additional datasets, EXIST-2022 and TRAC-2020, which help with the sexism detection task. Our approach focuses on Pre-Trained Models (PTM) architecture, combines using pre-trained language model Bert and BiLSTM network, which is inspired by Rodrıguez (Rodríguez-Sánchez et al., 2020) and Parikh (Parikh et al., 2019).

3 System overview

An overview of the architecture of our proposed model is depicted in Figure 2. The model we propose comprises the following components: Bert is used to extracting the embedding sequence of the word sequence. At the same time, the embedding sequence output by Bert is further feature extracted through BiLSTM to obtain context semantic information, and then sent to the output layer. The output layer is composed of a fully connected layer and a softmax layer, in which the dimension of the feature vector is adjusted in the fully connected layer and the softmax classifier is used to classify to achieve sexism detection.

3.1 Pre-trained Language Models

To better utilize the performance of large-scale pre-trained models, we employ the *pretrain-then-finetune* paradigm. The fine-tuning process of BERT involves utilizing the parameters obtained from pre-training as the initial values of the model while transferring the manually labeled dataset specific according to the downstream tasks. By balancing the relationship between the data and the model, this process results in the acquisition of a downstream-ready model (Devlin et al., 2018; Li et al., 2022).

As illustrated in Figure 2. In the BERT model, C refers to the pooled output of the first classification token ([CLS]), which summarizes the entire input sequence and is appropriate for sentence-level tasks like text classification and natural language inference. In contrast, T_i, which refers to the output representation of the i-th token in a sequence excluding the special token [CLS], is valuable for performing token-level tasks such as sequence labeling and question answering. Moreover, the "all_hidden_states" is a list that contains the hidden states of all layers of the model. Each hidden state is represented as a tensor with the shape (batch_size, sequence_length, hidden_size), where each state represents the hidden state for every token at each time step. Specifically, "all_hidden_states[0]" corresponds to the initial hidden states of the input sequence, while "all_hidden_states[10]" corresponds to the hidden states of the 11th layer. The inputs for the BiLSTM architecture were obtained using Bert, the "all_hidden_states [10]", which provides a more

detailed and contextualized view of the input sentence by capturing the interdependence between words and their surrounding context, and enhances our understanding of the input sentence without significantly increasing the model complexity.

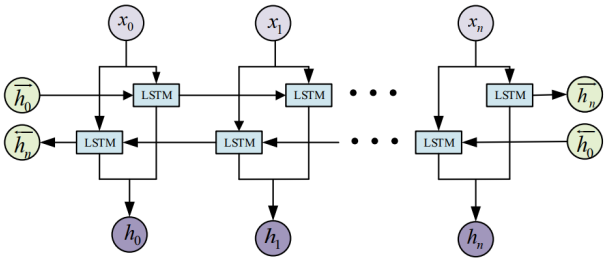

Figure 3: The structure of BiLSTM.

3.2 BiLSTM Model

The BiLSTM(Cho et al., 2014) model stands for Bidirectional Long Short-Term Memory, which is a model composed of a forward LSTM and a backward LSTM, and the LSTM in this model is a variant of the recurrent neural network. BiLSTM processes the input sequence in both forward and backward directions, which allows it to capture both past and future contexts, and can capture longer-term dependencies in the input sequence by using a memory cell and multiple gating mechanisms, thereby enabling it to be more robust to input noise. The structure of BiLSTM is shown in Figure 3.

To address the issue of irregular feature patterns, we utilize the BiLSTM model to uncover more hidden features in the contextual and semantic dependencies (Li et al., 2022). Specifically, we train BiLSTM on top of the BERT layer output to improve feature fitting and generalization performance for detecting sexism datasets.

4 Experimental

4.1 Task Description

Task A is a binary classification task for text: given a text sample from Gab and Reddit, the model is required to predict whether it contains sexism or not. The following examples present comments or statements that are either sexist or not sexist, respectively.

(1) `The world is filled with cunty women. Take solace in the fact that most of them will die bitter and alone with their 10 cats.....who will eat their faces off when they're dead. (#sexist)`

Splits	Class Label	Instances	total
train	sexist	3,398 (24%)	14,000
	non-sexist	10,602 (76%)	
dev	sexist	486 (24%)	2,000
	non-sexist	1,514 (76%)	
test	sexist	970 (24%)	4,000
	non-sexist	3,030 (76%)	

Table 1: Distribution of class labels of EDOS dataset.

(2) `This is why I always get suspicious when women comment about video games. I have a feeling most women think this about games. (#not sexist)`

Example 1 is a sexist text where this sentence expresses abuse and discrimination against women.

4.2 Evaluation Metrics

For all tasks (A, B and C), standard evaluation metrics that Macro F1 score is used to evaluate the participating system performance. Macro F1, a commonly used evaluation metric in machine learning, calculates an equal-weighted average of F1 scores for each class in a classification task and is widely adopted in the field of imbalanced classification tasks. (Han et al., 2022). calculated as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4)$$

where the variables TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

4.3 Dataset

Table 1 provides a summary of the detailed statistics for the EDOS dataset [1]. As shown in Table 1, the training data exhibits an imbalance, with 3398 positive samples and 10602 negative samples. Our pre-processing step only involves the removal of excess spaces, tabs, and line breaks.

[1] https://codalab.lisn.upsaclay.fr/competitions/7124

Dataset	Class Label	Instances
EXIST 2022	sexist	3377 (48.4%)
	non-sexist	3600 (51.6%)
TRAC-2020	GEN	382 (7.2%)
	NGEN	4947 (92.8%)

Table 2: Distribution of class labels in extra dataset.

4.4 Experiment Details

Due to the long-tail nature of training data by sexist discrimination, we adopted re-sampling and re-weighting techniques to address the issue, which did not have a significant effect. Therefore, we introduced two additional datasets for data augmentation, EXIST 2022 [2] and TRAC-2020 [3], for the same task of sexism detection. As shown in Table 2, the detailed statistics of these two datasets are summarized.

The method used for task A on sexism detection involves two main steps, namely, the extraction of embedding features and the semantic feature extraction step. Before the two main steps, We merge the three datasets and unify the labels, then the text for each instance is preprocessed by removing extra spaces, tabs, line breaks and URLs. In the embedding feature extraction step, each text instance is transformed into a 768-dimensional embedding vector by using the pre-trained Bert embedding model (Devlin et al., 2018). In particular, we used the base uncased model of Bert, which consists of 12 layers alongside 768 units per layer. In the semantic feature extraction step, we tried multiple neural network models including CNN, LSTM, DPCNN and BiLSTM. After a lot of experimental comparisons, we choose BiLSTM for feature extraction to obtain context semantic features.

During training, We did not perform pre-training on the BERT model and only fine-tuned it on downstream tasks by incorporating dataset labels. We use the Adam optimizer (Kingma and Ba, 2014) and the learning rate is 5e-5. We also use dropout (Hinton et al., 2012) with a rate of 0.3 to prevent overfitting. In particular, we use Focal Loss (Lin et al., 2017) as the loss function, alpha is set to 0.25, and gamma is set to 1. We conduct all the experiments on a machine equipped with a CPU: Intel(R) Xeon(R) Gold 6139 CPU @ 2.30GHz, and 6 GPU: Nvidia GeForce RTX 3090 GPU.

[2] http://nlp.uned.es/exist2022/

[3] https://sites.google.com/view/trac2/

Dataset	Model	Performance
EDOS 2023	Bert + BiLSTM	75.86%(∗)
EDOS 2023	Bert + LSTM	75.65%
	Bert + DPCNN	77.40%
EXIST 2022	Bert + BiLSTM	78.95%(∗)
TRAC-2020		

Table 3: The official test-set performance of Task A under different experimental settings. Scores marked with an asterisk denote the final submitted results.

4.5 Results

Table 3 shows a comparison of the final performance on task A official test set using the proposed model settings. Due to time constraints, we only used the monolingual pre-trained base uncased model of BERT to submit the fine-tuning results for Task A, achieving 78.95% F1 score. It is evident that BiLSTM outperforms other architectures at sexism detection task as it focuses on feature extraction, particularly in obtaining contextual semantic features. These tricks include introducing additional datasets and using different loss functions, which have helped to improve the model's performance to a certain extent.

It is worth noting that additional datasets were included for training, and thousands of Spanish language corpora were inadvertently introduced due to oversights in dataset preprocessing and merging. The reason why the model's performance has only marginally improved after data augmentation could be attributed to this factor. Specifically, the additional data set is mixed with about 20% of the Spanish language corpora. The performance of the model has been improved to a certain extent after removing it.

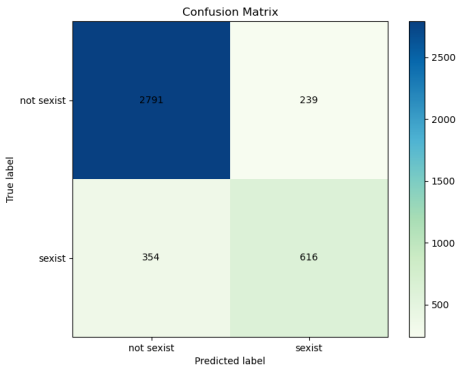

Figure 4: Confusion Matrix.

4.6 Further Analysis

The dataset is very unevenly distributed across sexist and non-sexist texts, which can cause the model to be highly biased towards non-sexist aspects. Therefore, we use the confusion matrix and segmentation metrics (including FP, TP, FN, TN) to show the difference of the model to the category. As shown in Fig. 4 and Table 4. Overall, our model performed relatively well on non-discriminatory text, with a precision of 89% and a recall of 92%, however, the model performed poorly on discriminatory text.

Label	Metrics	Performance
non-sexist	precision	88.74%
	recall	92.11%
	f1-score	90.40%
sexist	precision	72.05%
	recall	63.51%
	f1-score	67.51%

Table 4: Performance of segmentation metrics for each category.

5 Conclusion

In this paper, we introduce the system developed and evaluated for SemEval 2023 Task 10 and employ a novel neural framework that combines text representations obtained using pre-trained language model Bert, and using BiLSTM architecture to obtain context semantic information for better sexism detection. In addition, in order to alleviate the problem of data imbalance, we introduced additional datasets and tried different loss functions, which improved the performance of the model. Exploring approaches for detecting sexism in multilingual settings shows promise as a potential direction for future research.

Acknowledgement

We appreciate previous work (Abburi et al., 2021; Subies, 2021; Jiang et al., 2022; Parikh et al., 2021; Li et al., 2022) and open resources provided by Shao and Wu. Based on this, we conducted our work on EDOS task. We also appreciate the help from the EDOS team and program chairs. In addition to the traditional tools, we used Chatgpt to help polish the paper, just polish.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2021. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering*, 6(4):359–379.

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 141–153. Springer.

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.

Angelo Basile and Chiara Rubagotti. 2018. Crotonemilano for ami at evalita2018. a performant, cross-lingual misogyny detection system. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:206.

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.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Davide Buscaldi. 2018. Tweetaneuse@ ami evalita2018: Character-based models for the automatic misogyny identification task. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:214.

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.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder

for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Kevin Michael Deluca, Elizabeth Brunner, and Ye Sun. 2016. Weibo, wechat, and the transformative events of environmental activism on china's wild public screens. *International Journal of Communication (19328036)*, 10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. 2022. X-pudu at semeval-2022 task 6: Multilingual learning for english and arabic sarcasm detection. *arXiv preprint arXiv:2211.16883*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

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.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Xinlu Li, Yuanyuan Lei, and Shengwei Ji. 2022. Bert- and bilstm-based sentiment analysis of online chinese buzzwords. *Future Internet*, 14(11):332.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

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*.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

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, 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 de Lenguaje Natural*, 69:229–240.

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 *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584.

Guillem García Subies. 2021. Exist2021: Detecting sexism with transformers and translation-augmented data. In *IberLEF@ SEPLN*, pages 395–401.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.