Interpretable Sexism Detection with Explainable Transformers

Shamima Rayhana[1], Md Shajalal[2], Md Atabuzzaman[3] and Gunnar Stevens[2,4,*]

[1]Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur, Bangladesh
[2]University of Siegen, Siegen, Germany
[3]Virginia Tech, USA
[4]Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany

Abstract

With the widespread growth of social media platforms, instances of racism, cyberbullying, and the use of offensive language have surged. Consequently, women face challenges stemming from the presence of sexist content, which not only impedes their self-improvement but also exacerbates feelings of anxiety. Recognizing online sexism as a harmful phenomenon, the need for an automated tool to detect it has become paramount. This paper proposes an automated framework for extracting insights and identifying sexist language with high accuracy, utilizing machine learning (ML), deep learning (DL), and transformer-based models. Then, we incorporate the explainable AI (XAI) technique to enhance interpretability and make it more understandable to humans. To assess the performance of our method, we conducted experiments using a publicly available dataset focused on sexism. The experimental results underscore the effectiveness of our approach in detecting online sexism, surpassing the performance of several state-of-the-art methods.

Keywords

Sexism Detection, Explainability, LIME, Transformers, RoBERTa, XLM-R

1. Introduction

The Internet's pervasiveness in our daily lives stems from its vast array of essential services and resources. Its global reach bridges the gap between people from all corners of the world, fostering the growth of communities and driving progress in various facets of life. No other medium allowed every participant to communicate instantly with such a large audience [1].

Despite the numerous advancements in the Internet realm, a dark side exists associated with its usage. The Internet is home to potentially disturbing and harmful content showing negative behaviour, which can have significantly different impacts on different users [2]. While platforms like YouTube, Facebook, and Twitter have enabled information sharing and community building, they've also morphed into battlegrounds where people are bullied, smeared, and pushed to the margins simply for who they are [2, 3]. Particularly for women, the Internet often fosters a hostile environment for them, with widespread issues like racism, cyberbullying, body-shaming, and gender-based discrimination, especially in professional settings and on social media. The content produced by individuals expressing hatred tends to disseminate more rapidly, cover greater distances, and reach a considerably broader audience compared to the content generated by regular users [4]. These challenges instill fear and anxiety, deter women from pursuing their goals, harm self-esteem and cognitive abilities, creating an unwelcoming and harmful workplace.

The rapid spread of information on the internet, particularly within social networks, has heightened the severity of these harassing behaviours. Consequently, there is a pressing need for practical solutions to mitigate the harm inflicted by malicious propaganda in the digital realm [5]. Creating inclusive, secure platforms for women is crucial for effective human resource use, driving the advancement of NLP-based sexism detection methods.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.
md.shajalal@uni-siegen.de (M. Shajalal)

Identifying different forms of hate speech, such as racism, aggression, and misogyny, in social media poses a complex challenge [6]. This complexity arises from the varied meanings of slurs and offences, which depend on context, the gender and type of users, and how they are presented [7]. Moreover, the high volume of sexist posts, vocabulary discrepancies, and a hostile online environment make detection and moderation even more challenging, hindering efforts to create a safer space for women. Moderating user-generated content in offensive language identification for under-resourced languages is critical. It is feasible to recognize the key characteristics that aid in the identification of abusive content in the form of racist and sexist remarks that are a common occurrence on social media [8].

Due to advancements in natural language processing (NLP) technology, extensive research has been conducted on the automated detection of hate speech in textual content in recent years. Hate speech on social media is surging, prompting research into solutions and calls for stricter comment filtering by platforms [9]. GermEval-2018, SemEval-2019, and 2020 have advanced hate speech detection by organizing events and compiling diverse datasets. However, automated text-based detection has limitations. An explanation feature is crucial for filtering sexist content, ensuring transparency and trust. Many classifiers lack explainability, reducing confidence among users and moderators.

In this paper, we proposed methods to detect and explain sexist content. We sequentially used classical ML, deep learning (LSTM, BiLSTM, and CNN-BiLSTM), and transformers (XLM-R, RoBERTa, XLNet, and GPT2). Finally, we used the explainable AI (XAI) technique, LIME, to interpret model decisions. Experiments were conducted on the SemEval-2023 Task 10 dataset (Explainable Detection of Online Sexism) [10], and the results demonstrate the effectiveness of our approach in detecting sexist content with explanations. Our contributions to this paper are as follows:

- We conducted a wide range of experiments to detect sexist content utilizing ML, DL, and transformer-based approaches.
- We employed the XAI model to make the prediction and working principles of different approaches to humans understandable.

The subsequent sections of the paper are organized as follows: Section 2 is dedicated to examining pertinent literature in the field of sexism detection. Then, we present our methodology in section 3. In section 4, we thoroughly analyze the outcomes of our experiments. Finally, section 5 serves as both the conclusion for our methods and a platform for outlining our plans.

2. Related Work

Recently, there has been a growing interest in developing models to classify text, especially for hate speech detection. Across countries, laws against hate speech differ, but they typically target harmful expressions based on personal characteristics like race, color, national origin, sex, disability, religion, or sexual orientation [11]. This includes tasks such as automatically identifying sexist content on social media [5].

Automated detection of hate speech through machine learning is a promising field, as evidenced by the lack of review articles summarizing the available techniques [12]. For developing simplified methods for automatic detection of online sexism, conventional machine learning techniques are used. Waseem et al. [8] developed a machine learning model to automatically categorize a collection of tweets (16,000) based on the presence of harmful language patterns. Panwar et al. [13] encountered difficulty in identifying sexism using traditional methods as opposed to more recent approaches. They employed SVM, LR, RF, DT, XGBoost, and among these methods, RF demonstrated superior performance.

The application of neural network architectures, including CNNs and LSTMs, to detect hate speech within natural language text has gained significant traction in recent times. Badjatiya et al. [14] investigated the application of deep learning methods for the task of hate speech detection. They defined feature spaces for hate speech classifiers through task-specific embeddings generated using three distinct deep learning architectures: FastText, Convolutional Neural Networks (CNNs), and Long Short-Term Memory Networks (LSTMs). These models demonstrated impressive accuracy across

diverse NLP tasks. Zimmerman et al. [15] recommend focusing on hate speech detection and enhancing evaluation consistency in text classification, providing valuable guidance for the development of deep learning approaches.

Neural models along with transfer learning techniques lead to a significant improvement in performance across various text classification tasks, including the detection of hate speech. Language models based on transformers are becoming more and more proficient in natural language processing (NLP) assignments [16]. Mozafari et al. [17] explored the ability of BERT to grasp hateful context in social media content by applying novel fine-tuning methods rooted in transfer learning. In prior work on hate speech detection in Twitter data, DistilBERT was evaluated against attention-based recurrent neural networks and other transformer models, demonstrating superior performance and efficient parallelization compared to the baseline methods [18]. DistilBERT is a condensed iteration of the BERT model, providing a more lightweight and expedient alternative while still preserving comparable performance, and it is a good choice for tasks where computational resources are limited [19]. Bhatia et al. [20] used another transformer XLM-RoBERTa (XLM-R) to identify hate speech. They first applied Emoji2Vec to understand the meaning of emojis, then created word embeddings for hashtags. Finally, they combined these three resulting representations before classifying the text.

SamEval organized a competition to detect online sexism [10]. Among many submissions, multitask learning achieved the first rank in online sexism content identification using 'RoBERTa-large' and 'DeBERTa-V3-large', attaining a test F1-score of 0.8746 on sub-task A [21]. Sorensen et al. [22] claimed that using prompt-tuning-based ensemble F1-score for task A was greater than the first score of the competition, in which they primarily used a BERT-based ensemble. Ensemble-based transformer models came out with a better solution for this task, finding the ensemble size for best macro F1 [23, 24]. Pan et al. [25] proposed a new approach, Encoder + GCN + Adversarial training (their own) came up with the solution of sexism detection by representing features using the BERTweet model, a transformer, and a bidirectional LSTM layer. Though some approaches performed comparatively better than ours, their models did not explain why a sentence is sexist or not-sexist. All the models worked in a black-box manner. As a result, humans are unable to understand why a text is sexist or not-sexist. Therefore, in this paper, we present similar models with interpretation so that the models work in a human-understandable way.

3. Method

In this section, we detail our proposed method, consisting of pre-processing data to unlock the hidden gems within the text and different types of ML, DL, and Transformer-based models to extract meaningful insights and accurately identify sexist language. It also allowed us to capture the multifaceted nature of online sexism and identify even the most subtle instances of biased language.

3.1. Detection Models

This study uses traditional ML, DL, and transformer-based models to detect sexist content. We employed unigrams, bi-grams, TF-IDF, and word embeddings with ML to handle different data structures and capture key features by unlocking the hidden meanings within the text. We used deep learning techniques such as LSTM, BiLSTM, and a hybrid CNN-BiLSTM to capture contextual nuances and patterns within targeting sexist content. Transformer-based models like BERT, DistilBERT, XLM-R, RoBERTa, XLNet, and GPT-2 were also considered for effectively identifying sexist content, employing a locally explainable model for classification.

Classical Machine Learning Models: By analyzing the dataset, the machine learning algorithm builds a model to identify linguistic patterns in sexist text, improving classification accuracy. We applied several classical models, including logistic regression (LR), support vector machine (SVM), XGBoost, and Random Forest (RF), each chosen for its strengths. Since text data was unstructured, feature engineering was necessary to extract key characteristics in a structured format. To capture syntactic and semantic

aspects, we used unigrams, bi-grams, TF-IDF, and word embedding, enabling effective detection of online sexism.

Deep Learning Models: We employed three deep learning models—LSTM, BiLSTM, and CNN-BiLSTM—for detecting sexist content. LSTMs handle sequential dependencies, BiLSTMs enhance contextual understanding by processing text bidirectionally, and CNN-BiLSTM combines feature extraction with long-range dependency capture. These models effectively identify sexist content by leveraging advanced NLP techniques.

Transformer-based Models We also considered BERT, DistilBERT, XLM-R, RoBERTa, XLNet, and GPT-2 for this study. For BERT and DistilBERT, we employed the 'bert-base-uncased' and 'distilbert-base-uncased' models. For RoBERTa, we used the 'cardiffnlp/twitter-roberta-base-hate' model, and for XLM-R, the 'xlm-roberta-base' model was used. All transformer models were sourced from Hugging Face [1]. For BERT, DistilBERT, XLM-R, and RoBERTa, we used a learning rate of 1e-5, AdamW as the optimizer, and a batch size of 32. For XLNet and GPT-2, the learning rates were 2e-5 and 5e-5, respectively, with a batch size of 8 for both. Finally, we employed a locally explainable model to classify sexist texts.

4. Experiments

4.1. Dataset

For our experiments, we used the dataset provided by SemEval-2023 Task 10 [10]. This dataset contains 20,000 social media posts from Gab and Reddit [10], which are labeled as either sexist or not-sexist. Half of the data is from Reddit, and the other half is from Gab. The ratio of not-sexist to sexist samples is 3:1. The dataset is split into three parts: 14,000 samples for training, 2,000 samples for validation, and 4,000 samples for testing. Only the training dataset contains human-annotated labels and is publicly available. Therefore, we used the training dataset for our experiments (65% for training, 10% for validation, and 25% for testing purposes).

4.2. Experimental Setup

We conducted experiments with different settings to test how well our models perform in detecting online sexist content.

(i) Dataset Collection and Pre-processing: We collected the dataset from SemEval-2023 Task 10 [10], we prepared the data by removing noise. Preprocessing was performed according to the dataset structure.

(ii) Feature Extraction: In the subsequent analysis stage, known as Feature Extraction, relevant features were derived from textual inputs to transform unstructured text sequences into structured features. We used several techniques for feature extraction, these are count-vectorization, Word-embedding(fasttext), as well as how often words appear in a document (TF-IDF). Words were converted into numbers using vectorization by measuring the significance of a term within a sentence or corpus. This process considered both individual words (unigrams) and pairs of words (bigrams). The resulting vectors were then used to train machine learning models to predict whether a post is sexist.

(iii) Model Training: Machine learning, deep learning, and transformer-based models were trained on the dataset, split into 65% training, 10% validation, and 25% testing. For machine learning classifiers (Logistic Regression, SVM, XGBoost, and Random Forest), a 75%-25% split was used, with under-sampling and oversampling applied to training data to address imbalance. Deep learning models were trained using a batch size of 128 and the Adam optimizer. Transformer-based models, including BERT, DistilBERT, RoBERTa, XLNet, and GPT-2, were also trained. The models produced binary outputs for Task A (sexist vs. not-sexist) and multi-class outputs for Task B, classifying instances into four categories of sexist content: threats, derogation, animosity, and prejudiced discussion.

[1] https://huggingface.co

Table 1
Performance of Word Embedding on online sexism detection.

Methods	Techniques	Models	Accuracy	F1 Score
Word Embedding	fasttext / Word Embedding	XGBoost	0.7707	0.6186
		SVM	0.7657	0.5577
		LogisticRegression	0.7521	0.4918
		Random Forest	0.7510	0.4554
	fasttext + Under Sampling	XGBoost	0.7005	0.6758
		SVM	0.7051	0.6658
		Random Forest	0.6850	0.6173
		LogisticRegression	0.6735	0.5935

(iv) Evaluation: Performance evaluation metrics quantify the discrepancies between actual and predicted values through mathematical measures. In this final stage of the classification pipeline, model effectiveness was assessed using accuracy and F1-score.

4.3. Experimental Results

In table 1, when FastText was used for word embedding, the highest F1-score of 0.6758 was achieved by XGBoost. Among the machine learning methods, Random Forest demonstrated the best performance with the highest accuracy of 0.835.

Based on the experimental results of the baseline methods, the least effective performance among classical machine learning algorithms was observed when using bigram features. Bigrams, which capture pairs of adjacent words, may not adequately represent the complex linguistic patterns and nuances often present in texts related to sexism. As a result, this approach may struggle to capture the deeper semantic relationships necessary to accurately identify sexist content.

Table 2 highlights the effectiveness of deep learning approaches. In particular, the BiLSTM model achieved the highest accuracy of 0.8319, while both BiLSTM and LSTM yielded an identical F1-score of 0.74. Table 3 shows that the transformer-based GPT-2 model attained the highest accuracy and F1-score, with values of 0.854 and 0.8476, respectively. Compared to the top-performing system in SemEval-2023 (F1-score: 0.8746 on Task A), our GPT-2-based approach achieved a competitive F1-score of 0.8476 while also providing explainability through LIME—an aspect not addressed by the top system [21]. Additionally, a slightly higher F1-score of 0.8495 was reported using RoBERTa-large with domain-specific pretraining [25]. Overall, our methodology achieved near state-of-the-art performance and outperformed a semi-supervised multi-task system (F1-score: 0.8225) [26].

While many existing models prioritize performance alone, our approach combines competitive results with explainability, providing transparent insights into the rationale behind the classification of content as sexist.

Table 2
Performance Deep Learning model on online sexism detection.

Methods	Techniques	Accuracy	F1 Score
DL	LSTM	0.8132	0.74
	BiLSTM	0.8319	0.74
	CNN-BiLSTM	0.8118	0.73

4.3.1. Task-B

Table 4 depicts the results of Task-B. Classical machine learning algorithms were applied to detect four categories of sexist content (threats, derogation, animosity, and prejudiced discussion). When employing unigram techniques, Random Forest achieved the highest accuracy of 0.792, while SVM recorded the lowest accuracy at 0.7368. However, in terms of F1-score, SVM outperformed the others with the highest score of 0.3388, whereas Random Forest yielded the lowest F1-score of 0.3146.

Table 3
Performance of Transformer-based model on online sexism detection.

Methods	Techniques	Models	Accuracy	F1 Score
Transformer	BERT	CNN	0.7751	0.76
	DistilBERT	CNN	0.7726	0.77
	XLM-R		0.8234	0.73
	RoBERTa		0.8118	0.72
	XLNet		0.834	0.8167
	GPT2		**0.854**	**0.8476**

Random Forest again demonstrated superior accuracy, reaching 0.7545 with bigram analysis. However, the F1-scores were less impressive with bigrams compared to unigrams. In Task B, the highest F1-score of 0.2126 was achieved by SVM when employing bigram features.

4.3.2. Task-A

Using TF-IDF, XGBoost achieved a maximum F1-score of 0.3328 with an accuracy of 0.7868. XGBoost also demonstrated commendable accuracy when employing FastText techniques. However, the overall F1-score remained sub-optimal, similar to the performance observed with bigram analysis.

A comprehensive analysis of the Task B results revealed that Random Forest achieved the highest accuracy across all feature extraction techniques. The unigram and TF-IDF approaches outperformed both bigram and FastText methods. Bigram models, which capture consecutive word pairs, appeared less effective in capturing the nuanced context of sexist language. Similarly, FastText embeddings may have failed to adequately represent the subtle distinctions in sexist content, leading to lower performance compared to TF-IDF. In contrast, TF-IDF, by explicitly representing the importance of words within the context of the entire document, consistently demonstrated superior performance.

Table 4
Task-B Performance of Baseline method on online sexism detection.

Methods	Techniques	Models	Accuracy	F1 Score
Baseline	Unigram	SVM	0.7368	**0.3388**
		LogisticRegression	0.7685	0.3355
		XGBoost	0.7871	0.3198
		Random Forest	**0.792**	0.3146
	Bigram	SVM	0.7505	0.2126
		Random Forest	0.7545	0.2054
		XGBoost	0.7525	0.1932
		LogisticRegression	0.7485	0.1768
	TF-IDF	XGBoost	0.7868	0.3328
		SVM	0.7734	0.3159
		Random Forest	0.7888	0.3027
		LogisticRegression	0.7734	0.2659
	Fasttext	XGBoost	0.7548	0.2113
		SVM	0.7534	0.2038
		LogisticRegression	0.7517	0.1893
		Random Forest	0.7505	0.1735

4.4. Explaining the Prediction

In this section, we discuss the explainability/interpretability of our proposed online sexism detection models. Though SemEval-2023 Task 10 was an explainable detection of online sexism, no participant provided any human-understandable explanation of their model's detection of sexism. Therefore, we employed LIME, a local explanation model. Figure 1 represents two texts predicted as sexist and not-sexist by the BiLSTM model, and the true class is also non-sexist. From figure 1, we can see that the

most weighted word is ‘computer’ for non-sexists and the most sexist word is ‘women’. Similarly, the presence of the word ’slut’ makes the text sexist. Human reading or seeing the predicted explanation will be able to understand why the text is non-sexist or sexist.

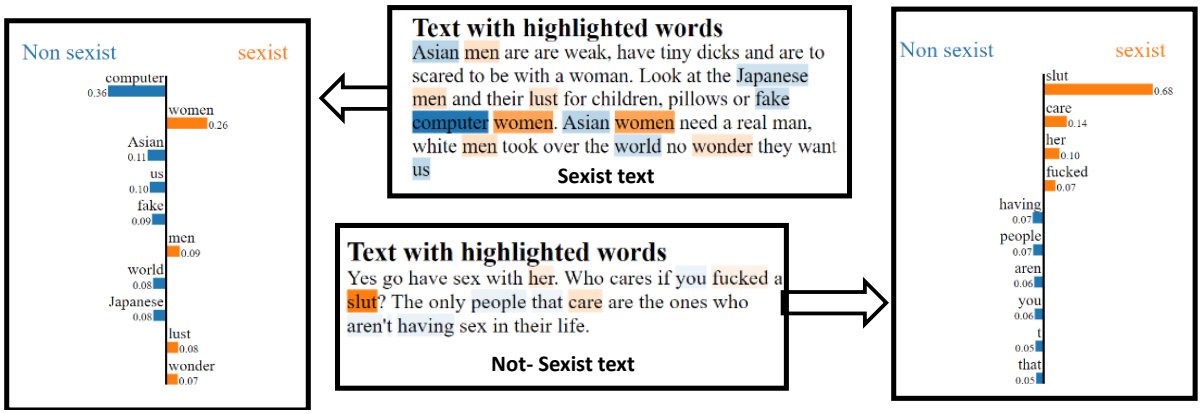

Figure 1: Explanation of sexist and not-sexist texts.

5. Conclusion and Future Work

This study aims to illuminate the detection of online sexism using three distinct categories: traditional machine learning (ML), deep learning (DL), and Transformer-based architectures. Although performance metrics varied between the models employed, with Random Forest (RF) achieving the highest accuracy and GPT-2 securing the best accuracy and macrof1 score, a key concern remains the inherent black-box nature of these models. To address this, we introduced an explainable technique to enhance human understanding of the detection process. Although we acknowledge the existence of more advanced methods that potentially exceed our proposed approach, we also highlight its ability to outperform some existing works. Future research endeavours will focus on diversifying the scope of investigated classes for further refinement and generalizability.

Acknowledgement

This research has been funded by the AntiScam Project (Defense against communication fraud), funded by BMBF Germany, Grant reference 16KIS2214

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] T. LaQuey, J. Ryer, et al., Internet companion, Addison Wesley Longman, 1994.
[2] T. Keipi, M. Näsi, A. Oksanen, P. Räsänen, Online hate and harmful content: Cross-national perspectives, Taylor & Francis, 2016.
[3] D. Benikova, M. Wojatzki, T. Zesch, What does this imply? examining the impact of implicitness on the perception of hate speech, in: Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings 27, Springer, 2018, pp. 171–179.
[4] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of hate speech in online social media, in: Proceedings of the 10th ACM conference on web science, 2019, pp. 173–182.
[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[6] F. Fasoli, A. Carnaghi, M. P. Paladino, Social acceptability of sexist derogatory and sexist objectifying slurs across contexts, Language sciences 52 (2015) 98–107.
[7] C. Hardaker, M. McGlashan, "real men don't hate women": Twitter rape threats and group identity, Journal of Pragmatics 91 (2016) 80–93.
[8] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[9] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[10] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, arXiv preprint arXiv:2303.04222 (2023).
[11] J. T. Nockleby, Hate speech in context: The case of verbal threats, Buff. L. Rev. 42 (1994) 653.
[12] A. Al-Hassan, H. Al-Dossari, Detection of hate speech in social networks: a survey on multilingual corpus, in: 6th international conference on computer science and information technology, volume 10, 2019, pp. 10–5121.
[13] J. Panwar, R. Mamidi, Panwarjayant at semeval-2023 task 10: Exploring the effectiveness of conventional machine learning techniques for online sexism detection, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1531–1536.
[14] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proc. of the 26th Int. conference on World Wide Web companion, 2017, pp. 759–760.
[15] S. Zimmerman, U. Kruschwitz, C. Fox, Improving hate speech detection with deep learning ensembles, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.
[16] S. Sai, N. D. Srivastava, Y. Sharma, Explorative application of fusion techniques for multimodal hate speech detection, SN Computer Science 3 (2022) 122.
[17] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8, Springer, 2020, pp. 928–940.
[18] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).
[19] H. Mohammadi, A. Giachanou, A. Bagheri, Towards robust online sexism detection: a multi-model approach with bert, xlm-roberta, and distilbert for exist 2023 tasks, Working Notes of CLEF (2023).
[20] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, A. Dash, One to rule them all: Towards joint indic language hate speech detection, arXiv:2109.13711 (2021).
[21] M. Zhou, Pinganlifeinsurance at semeval-2023 task 10: Using multi-task learning to better detect online sexism, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2188–2192.
[22] J. Sorensen, K. Korre, J. Pavlopoulos, K. Tomanek, N. Thain, L. Dixon, L. Laugier, Juage at semeval-2023 task 10: Parameter efficient classification, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1195–1203.
[23] A. Rydelek, D. Dementieva, G. Groh, Adamr at semeval-2023 task 10: Solving the class imbalance problem in sexism detection with ensemble learning, arXiv preprint arXiv:2305.08636 (2023).
[24] D. Obeidat, H. Nammas, M. Abdullah, et al., Just_one at semeval-2023 task 10: Explainable detection of online sexism (edos), in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 526–531.
[25] R. Pan, J. A. García-Díaz, S. M. Jiménez-Zafra, R. Valencia-García, Umuteam at semeval-2023 task 10: Fine-grained detection of sexism in english, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 589–594.
[26] S. Lamsiyah, A. El Mahdaouy, H. Alami, I. Berrada, C. Schommer, Ul\& um6p at semeval-2023 task 10: Semi-supervised multi-task learning for explainable detection of online sexism, in: The 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Toronto, Canada, Unknown/unspecified, 2023.