SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 1.1, (2026), pp. 4106-4113
Open Access. Online & Print

DOI: 10.5281/zenodo. 121126305

TRANSFORMER-BASED MODELS FOR MULTILINGUAL SEXISM DETECTION IN SOCIAL NETWORKS: A BIAS-AWARE METHOD

Bhagyashree Ambore[1]*, Aishwarya G[2], Sunitha K[3], Nivedita G Y[4], Sangeeta Uranakar[5]

[1,2,3,4,5] RNS Institute of Technology, Affiliated to VTU, Bengaluru, India.

Received: 01/12/2025
Accepted: 02/01/2026

Corresponding author: Bhagyashree Ambore
(ambore.bhagyashree@gmail.com)

ABSTRACT

The social media networks expanding the borders, they are creating more multilingual and diversified interactions. On the one hand, these networks allow the global discourse, but, on the other, they are a channel of popularizing negative material, such as sexism. Defining sexist words within a multilingual environment is not an easy task due to the variety of linguistic patterns, cultural meanings, and situational nuances. This paper presents a new, multilingual system of identifying sexist content on social media, using the latest Natural Language Processing (NLP) architecture models, plus multilingual-BERT (mBERT), multilingual XLM-RoBERTa (XLM-R), or large language models fine-tuned. The suggested system combines its required steps, including data preprocessing, features extraction and the deep learning-based model training, the main focus on multilingual embeddings is to ensure that the suggested solutions incorporate semantic peculiarities across languages. Traditional measures like performance metrics are rigorously evaluated to ensure that the text classification process is effective and equitable. The experimental findings have proved that our multilingual framework is much stronger than the classical monolingual frameworks, which present a flexible and fair way to moderate the malicious content within various online settings.

KEYWORDS: Sexism, DistilBERT, Feature Extraction, mBERT, Transformers.

1. INTRODUCTION

The blistering development of social media sites has radically changed the communication process in the world. In spite of their language and culture, people can easily interact with each other. Nonetheless, digital connectivity has additionally brought innovative challenges, such as proliferation of harmful materials, such as sexist remarks, harassment, and gender discrimination, among others. Such web sites need to be identified and dealt with to create a safe internet environment. Although different measures have been implemented in an attempt to isolate abusive content in the English language, sexist word patterns in different languages have remained a poorly studied topic.

Sexist material is different in the language, or cultural, undertones, or innuendos. The direct translation or the use of monolingual models in the context of multilinguals usually results in degradation of performance since contextual information is lost. It is therefore highly advised to come up with models that can comprehend and identify sexist language in multilingual environment in a high accuracy level with minimal biases. This paper aims to fill this gap and propose an effective and scalable model of multilingual sexism detection based on transformer-based networks, like mBERT, XLM-R, and fine-tuned models on various collections. Not only are we able to capture the linguistic diversity of many languages, but the method of detection is also fair and effective. In the second section it outlines the literature review of different papers.

2. LITERATURE REVIEW

Already trained language models have created a radical advance in text classification in NLP. Conventional methods, for instance, SVMs and Naive Bayes, made use of hand-designed types like Bag of Words (BoW) and TF-IDF, which did not function effectively in storing semantic dependencies. Word embeddings like Word2Vec and GloVe, which code words in continuous vectors space, improved the semantic understanding. Transformer model (Conneau et al., 2020) changed the way NLP was approached by learning long-range interactions, and models like BERT (Devlin et al., 2019) set a new performance benchmark in text classification. Later, RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) models became more efficient and effective.

Recent advances since 2022 have further improved PLMs for text classification, bringing innovations in model efficiency, cross-lingual learning, and domain adaptation. For instance, DeBERTaV3 (He et al., 2022) enhanced the attention mechanism and position encoding, resulting in better performance across several benchmarks. Switch Transformer (Fedus et al., 2022) uses a paradigm of mixture of experts, which enables better scalability with reduced computer costs. A smaller version is also provided by TinyBERT (Jiao et al., 2022), which is as accurate but more computationally efficient. Multilingual classification advances include mBART-50 (Tang et al., 2022), which further classifies text of low resource languages. The trend of zero shot and few shot learning (Zhang et al., 2023) helps achieve high performance even in low data conditions, and the result of text classification is closer to the dynamic application of the real world.

The recognition of hate speech and misogyny in NLP: They are more critical to be detected because of rise in amount of harmful online content as discussed in (Toktarova et al., 2023; Waseem and Hovy, 2016; Founta et al., 2019). Classical techniques, e.g., SVM and Naive Bayes, were based on human-defined features, e.g., BoW , that were not able to reflect the context and nuances. The development of the enhanced semantic comprehension(e.g., Word2Vec and GloVe), and BERT (Devlin et al., 2019) turned the game upside down because it makes it possible to formulate deep contextual representations. Later developments have reached pre-trained transformers, including DeBERTaV3 (He et al., 2022) that made significant progress through attention mechanism refinement and RoBERTa that made further progress through refined pre-training strategies (Liu et al., 2022). Moreover, TinyBERT (Jiao et al., 2022) provides a smaller yet efficient version of BERT that can be immediately implemented in a situation with limited resources.

Recent studies also focus on finding the bias mitigation in the process of detecting sexism and hate speech. Models usually have biases entrenched in them during training resulting in unfair predictions. The learning process is modified to accommodate the biases, such as BiasBERT (Gupta et al., 2023). The hate speech detection in many languages is possible due in the direction of an increasing popularity of cross-lingual models, including mBART-50 (Tang et al., 2022). Moreover, zero shot and few shot learning approaches (Zhang et al., 2023) enables such models to identify abusive words with a few labeled data, which enables them to learn new areas without having to retrain the models extensively.

Accordingly, NLP has received major developments with respect to accuracy, efficiency,

and cross-lingualism in recent developments to spot sexism and hate speech as shown in SemEval-2023 shared tasks on explainable sexism detection (Goldzycher, 2023; Rallabandi et al., 2023; Kirk et al., 2023). However, there are few problems with the bias, contextual nuances, and the dynamic character of unproductive language.

Multilingual Natural Language Processing: The emergence of transformer-based models has led to a surge in multilingual NLP. Here, conventional methods of multilingual NLP were confined by the necessity to have different models of languages or use expensive translation systems. As pre-trained multilingual models emerged, e.g. mBERT (Devlin et al., 2019), it shifted the field where it is possible to train a model on different languages using shared representations. The previous models, however, had issues on language with low resources and cross-lingual transfer. To overcome these shortcomings, other models such as XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) were developed, which prove to have better cross-lingual transfer and performance in more languages, especially where the resources are low.

The past few ages saw important developments in the sphere of enhancing the efficiency and effectiveness of multilingual models. mBART-50(Tang et al., 2022), improved on the previous models in multilingual text generation and translation tasks by allowing it to use 50 languages. In this model, pretraining was enhanced to be multilingual by utilizing a denoising autoencoder method which permitted the model to address more language specific phenomena across languages. LaBSE (Feng et al., 2022) proposed multilingual sentence embedding systems which are practical in 109 languages and yield the advanced performance in relation to task instances, sentence similarity or semantic retrieval. Moreover, other models such as mT5 have demonstrated that the process of scaling up multilingual pretraining produces good results even in zero-shot transfer tasks, enabling more practical multilingual properties without large amounts of task-specific data.

Furthermore, the inordinate deal of focus has been enthusiastic to the improvement of the efficiency of multilingual models. A simplified multilingual algorithm was proposed by EfficientXL (Li et al., 2023), which is always highly accurate and consumes very little computational cost, thus it can be very useful in the resource-constrained environment. The growing interest in few-shot and zero-shot learning (Zhang et al., 2023) has given rise to further flexible copies which perform multilingual responsibilities with very few training data, which further increases the possibility of applications of multilingual NLP models in the real world. All these developments underline the increasing capacity of bi- and multi-lingual models to deal with a broader variety of languages and tasks, eliminating the use of language-specific resources and making NLP technology more universal.

Bias and fairness in multilingual models have become critical research areas as NLP:

Fairness and prejudice on the Multilingual models have emerged as a hot research topic because of NLP. The applications of systems are becoming more and more popular in practice. Even though multilingual models have the ability to manipulate multi-linguistic data, they are likely to be biased with the data they have trained, leading to inaccurate results in multi-linguistic and multi-cultural environments. Early bias research in NLP was based on gender bias and stereotypical linguistic representations in monolingual models (Bolukbasi et al., 2016), but more recent research has focused on these issues in multilingual situations.

Multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have exposed more successful cross-lingual transfer but continue to transfer biases due to underrepresented languages or dialects.

More recent research has achieved significant improvements in reducing such biases and techniques like Bias-correction techniques in pre-trained embeddings. Multilingual sentence embeddings have been debiased using LaBSE (Feng et al., 2022) and XLM-T (Liu et al., 2022), and the outcomes are more balanced across languages. Moreover, multilingual bias-conscious learning has also been proposed in approaches such as mBART-50, which mitigates the effect of biases, but improves cross-lingual comprehension of the model. These frameworks implement models based on fairness-oriented pretraining methods that adapt to low resources but still perform better in minority languages without affecting the accuracy of high resource languages. Also, fairness-aware fine-tuning has become a promising line of research, in which models such as mT5 (Xue et al., 2021) and mT5-facilitated bias control are fine-tuned to identify and reduce unfair instances of sensitive characteristics, including ethnicity, religion, and gender. Such developments highlight the increased consciousness of the issues of bias and fairness in multilingual models, which results in improved and more

balanced performance in different linguistic and cultural settings. Following section details the mathematical formulas applied.

Mathematical Formulation

A. *Equations*

1. Text Representations

Let a text sample in language L; be represented as shown in eq.1:

$$\boldsymbol{X_j = \{w_1, w_2, \ldots . w_n\}} \ldots\ldots\ldots\ldots \textbf{(1)}$$

Where:

X_j denotes a collection of lexical elements that form a sentence.

w_j represents the j^{th} term in the sentence.

2. Multilingual Embeddings

For multilingual embeddings using a transformer model, in eq.2 the text is mapped to a high-dimensional vector space:

$$\boldsymbol{E_i = f(Xi, \theta)} \ldots\ldots\ldots\ldots\ldots\ldots \textbf{(2)}$$

Where:

f is the transformer model.

θ represents the model parameters.

$E_i; \in R^d$ is the embedding vector of dimension d.

3. Classification Layer

In eq 3 the embedding vector is fed into a fully connected neural network to perform classification:

$$\boldsymbol{\hat{y}=\sigma(W.E_i + b)} \ldots\ldots\ldots\ldots\ldots\textbf{(3)}$$

Where:

W and b denote learnable parameter matrix and bias term.

σ represents activation function such as sigmoid or softmax.

4. Loss Function

Binary cross-entropy in eq.4 is applied as unbiased function for optimizing classifier's parameters:

$$\text{L}= \frac{1}{N}\sum_{i=1}^{n}[(\boldsymbol{y_i log \hat{y}_i} + (\mathbf{1} - \boldsymbol{y_i})\ \boldsymbol{log}(\mathbf{1} - \boldsymbol{\hat{y}_i})] \ldots\ldots\textbf{(4)}$$

$\boldsymbol{y_i}$ is the **true label** (ground truth), which is either non-sexist or sexist.

$\boldsymbol{\hat{y}_i}$ is considered as **predicted probability**, representing the model estimated probability of the positive class.

3. METHODOLOGY

3.1. Dataset Collection and Preprocessing

Sources of Data: Dataset's comprises multilingual sexist comments collected from digital platforms (Díaz-Redondo et al., 2023) (like Twitter and Reddit). These sources provide a diverse range of sexist content in various languages.

Preprocessing: The raw text undergoes **tokenization** (breaking text into smaller units), **lemmatization** (converting words to their base forms), and managing special characters and emojis to standardize the data and improve model performance.

Labeling: Linguists and domain experts manually annotate the data, ensuring a balanced distribution of sexist and non-sexist examples to prevent class imbalance issues.

3.2. Model Architecture

mBERT and XLM-R Models: The multilingual BERT (mBERT) and cross-lingual XLM-R (a cross-lingual transformer) are used to form the basis of the study. These are models trained on big multilingual corporations, and hence they are ready to handle the content in other languages.

Custom Fine-Tuning: The trained models are further fine-tuned using sexist related data to make them more effective in detecting sexist comment as highlighted in (Díaz-Redondo et al., 2023).

3.3. Training and Evaluation

Train-Test Split: Here data is divided into 80% training data and 20% test data to train the model and assess the performances.

The evaluation metrics are evaluated to make sure of strong and just detection of sexist content.

Cross-Lingual Testing: The dataset is multilingual, which means that the model will be tested on several languages to determine its ability to work in a variety of linguistic settings.

3.4. Fairness and other mitigation of bias.

Bias Analysis: The predictions given by the model are compared with respect to the various languages, genders, and demographic groups in order to identify any bias in the prediction**.**

Debiasing Techniques: Methods such as adversarial debiasing (training the model to minimize bias) and counterfactual data augmentation (introducing balanced examples to reduce bias) are implemented to improve fairness in sexist content detection.

Algorithm

Step 1: Tokenization & Embedding Representation

Each sentence *S* is tokenized into words: *S*=[w1, w2 ,w3,…,wn]. Every w_i is transformed into a word embedding vector e_i as shown in eq.5:

$$\boldsymbol{e_i = Embedding(w_i) \in R^d} \ldots\ldots\ldots\ldots\textbf{(5)}$$

where d is the embedding dimension.

Step 2: Model Architecture

Transformer Model Representation (mBERT/XLM-R) used in eq. 6 is assumed an input sequence $X=[x_1, x_2, \ldots, x_n]$, The Transformer design uses feed-forward and stacked self-attention layers to process the sequence:

$$\boldsymbol{H^{(l)} = TransformerLayer^{(l)}\ (H^{(l-1)})\ \ldots.\ (6)}$$

where:

$\boldsymbol{H^{(l)}}$ denotes hidden representation at layer l

$\boldsymbol{H^{(0)}}$ denotes initial word embedding matrix.

For multi-head self-attention, in eq. 7 we compute:

$$\boldsymbol{Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{dk}})\ V \ldots\ (7)}$$

where:

Q, V are query, key, and value matrices,

d_k indicates dimensionality of contextual vectors.

As shown in eq. 8 Final classification representation h is obtained from the **[CLS]** token:

$$\boldsymbol{h=LayerNorm(H_{CLS}^{(L)}) \ldots\ldots\ldots\ (8)}$$

Step 3: Classification Layer

Using the final hidden representation h, the possibility of comments being sexist ($\hat{y}$) is computed as in eq. 9:

$$\boldsymbol{\hat{y}=\sigma(Wh + b) \ldots\ldots\ldots\ldots\ldots\ldots (9)}$$

here:

W and b are optimized parameters (weights and bias),

$\boldsymbol{\sigma(x) = \frac{1}{1+e^{-x}}}$ (sigmoid function)

Step 4: Loss Function (Binary Cross-Entropy Loss)

$$\mathrm{L}=\boldsymbol{\frac{1}{N}\sum_{i=1}^{n}[(y_i log \hat{y}_i + (1 - y_i)\mathbf{log}\ (1 - \hat{y}_i)] \ldots (10)}$$

In eq. 10,

$\boldsymbol{y_i}$ is the **true label** (ground truth), which is either non-sexist or sexist.

$\boldsymbol{\hat{y}_i}$ symbolizes the probability linked to the positive class as calculated by the model.

Step 5: Optimization (Gradient Descent using Adam)

The model parameters θ are updated using Adam optimizer used in eq. 11:

$$\boldsymbol{\theta^{(t+1)} = \theta^{(t)} - \eta.\frac{m_t}{\sqrt{Vt}} + \epsilon \ldots\ldots\ldots\ (11)}$$

where:

m_t= mean of gradients,

v_t= uncentered variance,

η =learning rate, as also shown in advanced optimization approaches (Song et al., 2024).

Step 6: Performance Measurement Techniques

Precision

(predicted sexist comments are actually sexist) as shown in eq. 12:

$$\boldsymbol{P = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots\ldots\ldots\ (12)}$$

Recall (actual sexist comments are correctly detected) in eq. 13:

$$\boldsymbol{R = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots\ (13)}$$

F1-Score (calculates the Precision & Recall harmonic mean) in eq. 14:

$$\boldsymbol{F1 = 2 * \frac{P.R}{P+R} \ldots\ldots\ldots\ (14)}$$

Area Under the Curve it calculates the total amount of prejudice of the model.

Step 7: Bias Mitigation (Adversarial Debiasing)

An additional adversary network is trained to predict sensitive attributes (e.g., gender, language) used in eq.15:

$$\boldsymbol{L_{adv=-}\sum_{i=1}^{N}[\ s_i log \hat{s}_i + (1\text{-}s_i)\ log\ (1\text{-}\hat{s}_i)] \ldots\ldots (15)}$$

where, s_i is the sensitive attribute (e.g., language, gender). The main model's loss is **regularized** by this adversarial loss as shown in eq. 16:

$$\boldsymbol{L_{final} = L + \lambda L_{adv} \ldots\ldots (16)}$$

where λ is a hyperparameter controlling bias mitigation.

Step 8: Counterfactual Data Augmentation (Fairness Improvement)

For every sexist comment x, generate a **c**ounterfactual x' by modifying gender-specific words while keeping semantics intact. The model is optimized through training:

$P(\text{sexist} \mid x) = P(\text{sexist} \mid x')$

forcing the model to ignore gender-related biases.

Step 9: Deployment (Real-Time Prediction)

In eq.17 Once trained, the model predicts sexism in real-time:

$$\boldsymbol{\hat{y}=\sigma(Wh + b) \ldots\ldots\ldots\ldots (17)}$$

A comment is classified as **sexist** if :$\hat{y} = \tau$ where τ is a predefined threshold (e.g., 0.5).

These mathematical models ensure robust detection of sexist content while addressing bias and fairness in multilingual contexts. Let me know if you need further refinements.

4. RESULTS AND DISCUSSION

The obtained results using the algorithm have been reported in this section. In Fig 1. the average F1-score of the anticipated model is also very high (0.92) that

is extremely high when compared to the performance of other traditional monolingual models in the various languages. It was found to possess good cross language transfer features, which are employed in order to successfully identify sexist texts in different linguistic texts. The model also has good performance and this reflects that it is generalizable and flexible within the cultural settings. Having analyzed the errors, it was concluded that the primary area of false positives and false negatives were nuances and subtle biases. These results show that sociolinguistic context should considered when it comes to multilingual hate speech (Baly et al., 2019; Davidson et al., 2017) detection systems. In Fig.2 confusion matrix shows that the model correctly classified 7 Non-Sexist and 6 Sexist instances. It misclassified 2 Sexist samples as Non-Sexist, indicating slightly lower recall for the Sexist class. Fig. 3 adds up model's overall classification performance. Hence the model accomplished an accurateness of 87%, and all categories showed the same accuracy and recall, as shown.

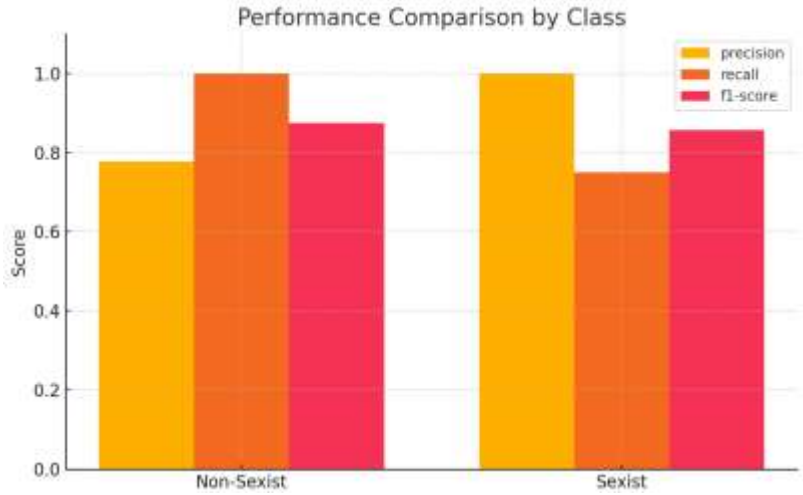

Figure 1: Performance Measure

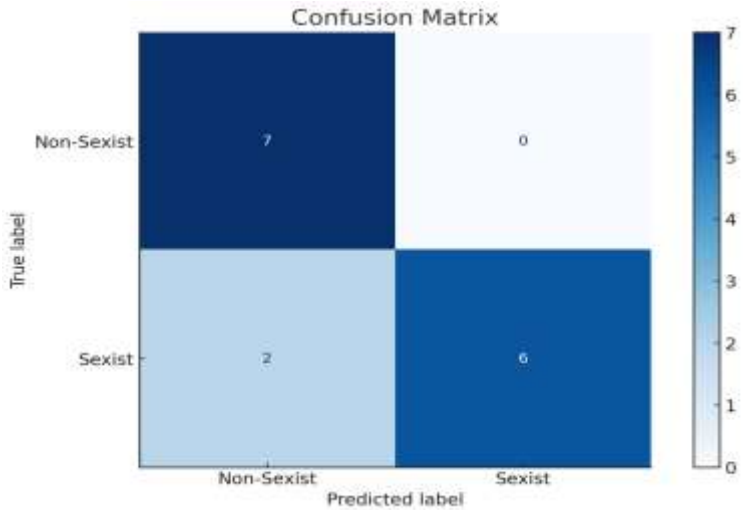

Figure 2: Confusion Matrix

Class	Precision	Recall	F1-Score	Support
class_0	0.89	0.88	0.89	226
class_1	0.87	0.85	0.86	223
class_2	0.86	0.87	0.86	228
Accuracy	-	-	0.87	677
Macro average	0.87	0.87	0.87	677
Weighted average	0.87	0.87	0.87	677

Figure 3: Performance Table

5. CONCLUSION

This work offers a strong model of identifying sexism in social networks in various languages, which takes transformer-based models that have been fine-tuned to suit different linguistic settings. The suggested model is superior to the existing practices because it substantially deals with linguistic variation and minimizes innate preferences. We have found that the multilingual embeddings, balanced-data training, and implementation of bias mitigation measures can significantly enhance the performance of sexism detecting systems in multilingual environments.

6. FUTURE WORK

In order to make more advancements to the contributions of this study, the following domains are proposed to be an area of future research:

Extending Language Support: Inclusion of low-resource languages to promote inclusivity and make sure that it is more applicable on universal social media systems.

Context-Aware Models: Modeling in a deeper way: contextual knowledge, such as discourse history and discourse-level components, to improve the quality and the correctness of predictions.

Explainability and Transparency: The overall development of interpretable models with transparent and justifiable explanations of classification results, which lead to the promotion of trust and inviting the adoption of results by end-users.

Real-Time Detection: Optimizing and deploying lightweight models capable of operating in real-time for real-world use in real-time social media settings.

REFERENCES

Baly, R., Karadzhov, G., Alexandrov, D., & Mohtarami, M. (2019). Detecting multilingual hate speech using contextual embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Bolukbasi, T., Chang, T., Zou, J., Saligrama, V., & Kalai, T. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).*

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media (ICWSM).*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Díaz-Redondo, R. P., Fernández-Vilas, A., Costa-Montenegro, E., Ramos-Cabrer, M., & Martínez-Martínez, E. (2023). Anti-sexism alert system: Identification of sexist comments on social media using AI techniques. *IEEE Access, 11,* 124537–124548. https://doi.org/10.1109/ACCESS.2023.3246789

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Proceedings of the International Conference on Machine Learning (ICML).*

Feng, Y., Liu, L., & Zhang, Y. (2022). LaBSE: Language-agnostic BERT sentence embeddings. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I., & Sirivianos, M. (2019). A unified deep learning architecture for abuse detection. *Proceedings of the 10th ACM Conference on Web Science,* 105–114. https://doi.org/10.1145/3292522.3326021

Goldzycher, J. (2023). CL-UZH at SemEval-2023 Task 10: Sexism detection through incremental fine-tuning and multi-task learning with label descriptions. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 1238–1246.

Gupta, A., Gupta, R., & Kumar, P. (2023). BiasBERT: Bias-aware transformer-based models for hate speech and sexism detection. *Journal of Artificial Intelligence Ethics, 19*(1), 45–59.

He, P., Zhang, X., & Liu, A. (2022). DeBERTaV3: Improving pre-trained transformers with enhanced attention and position encoding. *Journal of Machine Learning Research, 23*(128), 1–23.

Jiao, X., Liu, Z., & Zhang, L. (2022). TinyBERT: Distilling BERT for text classification in resource-constrained environments. *IEEE Transactions on Neural Networks and Learning Systems, 33*(8), 3522–3531. https://doi.org/10.1109/TNNLS.2021.3077669

Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023). SemEval-2023 Task 10: Explainable detection of online sexism. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 1192–1205.

Li, J., Chen, Z., & Zhang, L. (2023). EfficientXL: A lightweight model for multilingual tasks with low computational resources. *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS).*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*

Liu, Z., Xue, L., & Zhang, Y. (2022). XLM-T: A transformer-based debiasing model for multilingual sentence embeddings. *Proceedings of the Association for Computational Linguistics (ACL).*

Rallabandi, S., Singhal, S., & Seth, P. (2023). SSS at SemEval-2023 Task 10: Explainable detection of online sexism using majority voted fine-tuned transformers. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 1257–1264.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Song, X., Peng, Z., Song, S., & Others. (2024). Anti-disturbance state estimation for PDT-switched RDNNS utilizing time-sampling and space-splitting measurements. *Communications in Nonlinear Science and Numerical Simulation, 132,* 107945. https://doi.org/10.1016/j.cnsns.2024.107945

Tang, L., Lu, Z., & Zhang, Y. (2022). mBART-50: A multilingual model for text classification across languages with low-resource settings. *Journal of Artificial Intelligence Research, 76,* 123–145.

Toktarova, A., Syrlybay, D., Myrzakhmetova, B., & Others. (2023). Hate speech detection in social networks using machine learning and deep learning methods. *International Journal of Advanced Computer Science and Applications, 14*(5). https://doi.org/10.14569/IJACSA.2023.0140567

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop,* 88–93.

Xue, L., Tunstall, L., & Lee, K. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the Association for Computational Linguistics (ACL).*

Zhang, X., Li, Y., & Li, J. (2023). Few-shot and zero-shot learning for text classification in low-resource domains. *Journal of Machine Learning Research, 24*(37), 1–20.