Sexism Identification in Social Networks: Advances in Automated Detection - A Report on the Exist Task at CLEF

Nimra Maqbool[1,†]

[1]Information Technology University (ITU), Lahore, Pakistan.

Abstract

The widespread usage of social networks has brought forth numerous challenges, including the proliferation of sexist content, which contributes to gender inequality and discrimination. The Sexism Identification in Social Networks (SIT) task at the Conference and Labs of the Evaluation Forum (CLEF) aims to promote research and development of automated methods for identifying and mitigating sexism online. This review work provides an overview of the SIT task, discusses the dataset used, highlights the approaches and techniques employed, presents the evaluation metrics utilized, and offers insights into the future directions for advancing sexism identification in social networks. This review work describes the organization, goals, and results of the sExism Identification in Social networks (EXIST) challenge. EXIST 2024 proposes two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English. During CLEF workshop, we investigate a broad range of models, including traditional machine learning methods, such as ensemble models and probability-based model like Random Forest, XGboost and deep learning architectures models, like BERT and Multilanguage models. The Experimental results show promising results in these areas and especially multilingual BERT model outperform among models.

Keywords

Sexism, Gender discrimination, Social networks, Social media analysis, Text classification, Hate speech detection, Natural language processing

1. Introduction

In the ever-evolving digital landscape, social networks have become an integral part of our daily lives. They provide us with platforms to connect, share ideas, and express ourselves. However, as these networks have expanded their reach, they have also become breeding grounds for various forms of discrimination and prejudice. Among these insidious phenomena, sexism, characterized by the marginalization and objectification of individuals based on their gender, remains a pervasive issue in online spaces.

Furthermore, Sexism in social networks presents a multifaceted challenge, as its manifestations range from subtle micro aggressions to overt harassment and abuse. Identifying and combating sexism in these virtual realms is crucial to fostering inclusive and equitable online environments. By doing so, we can not only promote gender equality but also cultivate safe and empowering digital spaces that resonate with users from diverse backgrounds.

The identification of sexism in social networks presents a formidable challenge due to the vast amounts of user-generated content, diverse linguistic expressions, and intricate nuances involved in language interpretation. Nevertheless, recent advances in artificial intelligence (AI) and natural language processing (NLP) have sparked renewed hope for combating this societal ill. By leveraging computational techniques and machine learning algorithms, researchers have begun to develop automated tools capable of detecting and analyzing instances of sexism within social network data. Within this study, I embark on an exploration of diverse machine learning and deep learning models, aiming to effectively classify labels pertaining to sexism identification, including both cross-grained labels and fine-grained labels.

The primary objective of this research endeavour revolves around three pivotal tasks, namely:

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

bsce21012@itu.edu.pk (N. Maqbool)

1.1. Task 1: Sexism Identification

In this task, the focus lies on binary classification. The underlying model is designed to discern whether a given tweet encompasses expressions or behaviors that exhibit sexism. This involves analyzing the content of the tweet to identify linguistic patterns, phrases, and contexts that may indicate sexist attitudes or discriminatory language.

1.2. Task 2: Source Intention

Once a message is identified as sexist, the subsequent task endeavours to categorize the message based on the author's intention. This classification sheds light on the role played by social networks in the generation and dissemination of sexist messages. The proposed task involves a ternary classification with the following categories:

1. Direct: The intention of the message is to be inherently sexist or to encourage sexist behaviour.
2. Reported: The intention is to report and share a sexist situation experienced by a woman, either in the first or third person.
3. Judgemental: The intention is to pass judgment, as the tweet describes sexist situations or behaviours with the aim of condemning them.

1.3. Task 3: Sexism Categorization

In this task, the objective is to further classify the tweets, specifically identifying the facets of women that are frequently targeted within social networks. This categorization will contribute to the development of policies aimed at combatting sexism. Each sexist tweet must be classified into one or more of the following categories:

1. Ideological and Inequality: The text discredits the feminist movement, denies the existence of gender inequality, or portrays men as victims of gender-based oppression.
2. Stereotyping and Dominance: The text perpetuates false notions about women, suggesting that they are more suited for certain roles (such as mother, wife, caregiver, affectionate, submissive, etc.) or unsuitable for certain tasks (such as driving, labour-intensive work, etc.). It may also claim male superiority over females.
3. Objectification: The text reduces women to objects, disregarding their dignity and personal qualities, or describes specific physical attributes that women must possess to conform to traditional gender roles.
4. Sexual Violence: The text contains sexual suggestions, requests for sexual favours, or engages in sexual harassment.
5. Misogyny and Non-Sexual Violence: The text expresses hatred and displays acts of violence towards women, which may not necessarily be of a sexual nature.

By comprehensively investigating these tasks, employing robust machine learning and deep learning models; we aim to enhance our understanding of sexism identification within social networks. This research will contribute to the development of more effective strategies for combatting sexism and fostering an inclusive and equitable online environment.

2. Related Work

Sexism in social networks has emerged as a significant concern, highlighting the need for effective identification methods to mitigate its harmful effects. This section provides an overview of the existing research on sexism identification in social networks, including various approaches, methodologies, and key findings.

Francisco Rodr [1] presenting a multilingual system based on pre-trained transformers and comparing single task to multi task learning to identify sexism in social networks. Similarly Rolfy Nixon Montufar Mercado [2] proposed a model to detect cyberbullying in Spanish-language social networks using sentiment analysis techniques such as bag of words, elimination of signs and numbers, tokenization, stemming, and a Bayesian classifier. Dina Eliezer and Brenda Major[3] examines whether group identification moderates the extent to which perceived in-group discrimination is threatening, as indexed by physiological and self-report measures. Women read and gave a speech summarizing an article describing sexism as prevalent or rare.

Furthermore, Simona Frenda [4] propose first-time approach to the automatic detection of misogyny and sexism against women using the same computational approach. Along with that they also investigation of linguistic analogies and differences between sexism and misogyny from a computational point of view. He also examination of the usefulness of stylistic and lexical features for hate speech online against women. Parikh and Pulkit [5] introduces a neural framework for multi-label classification of sexism and misogyny in texts, combining BERT with word embeddings, and outperforming existing baselines. Devadath [6] examines Twitter sentiment analysis, focusing on NLP methods, tools, and machine learning algorithms like Naive Bayes. By evaluating metrics such as F1 score and precision, it categorizes a million tweets into positive and negative sentiments, emphasizing ethical considerations and informed decision-making.

Similarly, Devi and Bali [7]introduces a refined classifier for identifying racist and sexist comments on Twitter using NLP and ML techniques. With XGBoost and word2vec, the model attained 69% accuracy and an F1 score of 0.690285, showcasing promising results amidst the pandemic-driven digital culture shift. Furthermore, De Paula, Angel Felipe Magnoss [8] introduces a system for identifying and classifying sexism in English and Spanish social media using multilingual BERT models and ensemble strategies. It outperformed baseline models, securing first place in the EXIST 2021 task with high accuracies and F1-scores.

The EXIST campaign, focusing on online sexism detection, featured three tasks at CLEF 2023: sexism identification, categorization, and source intention identification,[9][10] as detailed by Plaza et al. (2024). Adopting a "learning with disagreement" approach, it aimed to reconcile differing perspectives in labeling, fostering equitable system development. The campaign's third edition, presented at the CLEF 2024 conference, provided new test and training data to enhance the identification and characterization of sexism in social networks and memes. With 28 participating teams and 232 submissions, the initiative underscored the research community's commitment to mitigating the impact of offensive content on women's well-being and freedom of expression on social media.

3. Utilization of Existing Models

3.1. Machine Learning Models

Two machine learning models, Random Forest and XGBoost, were implemented using TF-IDF and BOW techniques for binary classification in Task 1 and multi-class classification in Task 2.

3.1.1. Random Forest

Random Forest, a popular ensemble learning technique, constructs multiple decision trees during training, using random subsets of data and features to reduce overfitting and enhance generalization. By combining predictions from these trees, it produces robust and accurate final predictions.

It was implemented using both TF-IDF and BOW techniques. Hyperparameter tuning was performed using RandomizedSearchCV to optimize the model, and it was subsequently trained using the best parameters obtained through this process. This approach was adopted to achieve the best possible results.

3.1.2. XGBoost

XGBoost, or Extreme Gradient Boosting, is a fast and powerful implementation of gradient boosting algorithms. It sequentially builds decision trees to correct errors from previous trees, optimizing a specified loss function using gradient descent. Its efficiency and enhancements like regularization make it a popular choice for high-accuracy predictions in competitions and real-world applications.

The XGBoost model was implemented with both TF-IDF and BOW techniques. Hyperparameter tuning was performed using RandomizedSearchCV to identify the optimal parameters, and the model was trained using these best parameters. This approach aimed to achieve the highest performance.

3.2. Multi-language Bert Model

The Multilingual BERT (Bidirectional Encoder Representations from Transformers) model is a variant of the BERT model that is trained on text from multiple languages. BERT, developed by Google, is a pre-trained deep learning model that has achieved state-of-the-art results in various natural language processing (NLP) tasks.

The Multilingual BERT model is designed to handle multilingual text by jointly training on a large corpus of data from different languages. It learns a shared representation for words across languages, allowing it to capture cross-lingual semantic similarities and transfer knowledge between languages. This means that the model can understand and generate representations for text in multiple languages without needing language-specific models.

The training process of the Multilingual BERT model involves two key steps: pre-training and fine-tuning. During pre-training, the model is trained on a massive amount of monolingual text from various languages. It learns to predict missing words in sentences using a masked language modelling objective and also learns to determine whether two sentences follow each other in the original text or not, which helps it capture contextual relationships.

Once pre-training is completed, the model is fine-tuned on specific downstream NLP tasks such as text classification, named entity recognition, or sentiment analysis. Fine-tuning involves training the model on a smaller task-specific dataset in a supervised manner, where the model's pre-trained knowledge is adapted to the specific task at hand. This process allows the Multilingual BERT model to generalize across languages and perform well on various NLP tasks in different languages.

The Multilingual BERT model's success lies in its ability to capture contextual information from large-scale pre-training on diverse multilingual data. By leveraging the shared representations learned during pre-training, the model can handle a wide range of languages, even those with limited labelled data, and exhibit strong performance on tasks like text classification, information retrieval, and machine translation across multiple languages.

4. Experiments

The experiment section presents the findings and outcomes of the study on sexism identification in social networks. This section aims to provide a comprehensive analysis of the performance and effectiveness of the proposed approaches and methodologies. The results are presented in terms of various evaluation metrics, including accuracy, precision, recall, and F1-score. The section begins by summarizing the dataset used for evaluation, including the number of tweets, the languages covered, and the annotation process.

4.1. Dataset

Sexism encompasses any form of discrimination or bias against women based on their gender. In this study, we collected a large dataset of tweets in both English and Spanish, consisting of over 8 million tweets. The data collection period spanned from September 1, 2021, to September 30, 2022. To ensure a balanced dataset, we removed seeds (initial keywords) with fewer than 60 associated tweets. Ultimately, we obtained 183 seeds for Spanish and 163 seeds for English. The EXIST 2024 Tweets Dataset contains more than 10,000 labeled tweets, both in English and Spanish. In particular, the training set contains 6,920 tweets, the development set contains 1,038 tweets and the test set contains 2,076 tweets. Distribution between both languages has been balanced.

To address terminology and temporal biases, we carefully selected subsets of tweets for training, development, and testing. For each seed, we randomly selected approximately 20 tweets within the period from September 1, 2021, to February 28, 2022, for the training set. We aimed to maintain a representative temporal distribution within each seed's tweets. Similarly, we selected 3 tweets per seed from May 1, 2022, to May 31, 2022, for the development set, and 6 tweets per seed from August 1, 2022, to September 30, 2022, for the test set. To avoid author bias, we included only one tweet per author in the final selection. Additionally, we removed tweets containing fewer than 5 words. Consequently, we obtained over 3,200 tweets per language for the training set, around 500 per language for the development set, and nearly 1,000 tweets per language for the test set.

During the annotation process, we considered potential sources of "label bias." Label bias may arise due to socio-demographic differences among annotators or when multiple correct labels or highly subjective decisions exist. To mitigate this bias, we took into account two demographic parameters: gender (MALE/FEMALE) and age (18-22 y.o./23-45 y.o./+46 y.o.). Each tweet was annotated by six crowdsourcing annotators selected through the Prolific app, following guidelines developed by two experts in gender issues.

Given the highly subjective nature of sexism identification and the challenge of interpreting natural language expressions in context, we adopted a learning with disagreements approach. This paradigm allows systems to learn from datasets where no definitive "gold" annotations are provided but instead incorporates information about the annotations from all six annotators, capturing the diversity of perspectives. By training directly from the data with disagreements, rather than relying on an aggregated label, we aim to incorporate the varying annotations per instance across the six different annotator strata.

4.2. Results

In this section, we present the findings and results obtained from our study. we employed various classification metrics to assess the performance of the sexism identification models. These metrics include accuracy, precision, recall, and F1-score, which provide a comprehensive evaluation of the models' ability to correctly classify tweets as sexist or non-sexist.

Table 1
Performance metrics of TASK 1: Sexism Identification on Evaluation Dataset.

Model	Precision	Recall	F1-Score	Accuracy
Random Forest (TF-IDF)	0.71	0.72	0.72	0.72
Random Forest (BOW)	0.73	0.72	0.71	0.72
XGboost (TF-IDF)	0.74	0.65	0.72	0.72
XGboost (BOW)	0.74	0.73	0.69	0.72
Multilingual Bert model	0.78	0.78	0.77	0.78

In table 1 the BERT model outperformed other models such as Random Forest and XGBoost in the task of binary classification of sexism in tweets, which involved both Spanish and English languages. BERT's success can be attributed to its advanced transformer architecture, pre-trained representations, multilingual capability, and ability to generate contextualized embeddings, allowing it to effectively capture complex linguistic patterns and understand context. In contrast, models like Random Forest and XGBoost with TF-IDF and BOW may have plateaued in performance due to their reliance on fixed-length feature representations and limitations in handling multilingual text.

Table 2
Performance metrics of TASK 2: Source Intention on Evaluation Dataset.

Model	Precision	Recall	F1-Score	Accuracy
Random Forest (TF-IDF)	0.35	0.37	0.33	0.58
Random Forest (BOW)	0.37	0.34	0.31	0.55
XGboost (TF-IDF)	0.56	0.26	0.21	0.46
XGboost (BOW)	0.67	0.25	0.20	0.47
Multilingual Bert model	0.64	0.34	0.36	0.46

In table 2 XGBoost and BERT outperformed the alternatives. XGBoost's strength lies in its ability to handle structured data and capture complex patterns through boosting, while BERT's deep contextual understanding of text allows it to excel in nuanced tasks. These models also generalize well to unseen data. In contrast, random forest models with TF-IDF and BOW underperformed, likely due to data imbalance, difficulty in handling high-dimensional sparse features, and lack of contextual understanding. Consequently, XGBoost and BERT demonstrated superior performance in our experiments.

In Task 2, the random forest model using TF-IDF struggled primarily due to data imbalance, high-dimensional sparse features, and a lack of contextual understanding. Imbalanced datasets skewed predictions towards the majority class, while TF-IDF's high-dimensional sparse vectors posed challenges for random forests not adept at handling such feature spaces. The model's inability to capture contextual relationships in text further undermined its performance. Solutions could involve resampling techniques to balance data classes, feature dimensionality reduction via PCA, or exploring alternative feature representations like word embeddings. Similarly, in Task 2 with Bag-of-Words (BOW), the random forest faced issues related to sparse feature handling and inadequate contextual comprehension. Improving performance might entail employing ensemble methods, incorporating more advanced feature engineering techniques, or integrating hybrid models that combine BOW with richer contextual information to better suit the task's requirements.

Table 3
Performance metrics of Multilingual Bert model for TASK 3: Source Intention on Evaluation Dataset.

Class Name	Precision	Recall	F1-Score	Accuracy
Stereotyping-Dominance	0.42	0.41	0.39	0.40
Objectification	0.49	0.36	0.33	0.39
Sexual-Violence	0.44	0.46	0.42	0.44
Misogyny-Non-Sexual	0.41	0.36	0.37	0.38
Ideological-Inequality	0.34	0.33	0.38	0.35
Uknnown	0.0	0.0	0.0	0.0

In table 3, involving multi-label classification of the evaluation dataset using the multilingual BERT model, the evaluation metrics for all classes were in the 40s, indicating suboptimal performance. This likely resulted from the complexity of multi-label classification, data imbalance, and difficulty in capturing nuanced label distinctions. To improve performance, strategies such as data augmentation,

balancing techniques, fine-tuning BERT, optimizing decision thresholds, model ensembling, and considering label dependencies can be applied. Additionally, post-processing methods like label smoothing can further refine accuracy.

Table 4
Performance metrics of TASK 1, TASK 2 and TASK 3 on test dataset

Task	Model	Language	ICM-Hard	ICM-Hard Norm	F1-YES
Task-1	Random Forest	Combine	0.28	0.64	0.69
	Random Forest	ES	0.25	0.62	0.71
	Random Forest	EN	0.29	0.64	0.66
	XGboost	ES	0.33	0.66	0.71
	XGboost	EN	0.34	0.67	0.68
Task-2	Random Forest	Combine	-0.47	0.34	0.29
	Random Forest	ES	-0.51	0.33	0.29
	Random Forest	EN	-0.42	0.35	0.30
	XGboost	ES	-0.54	0.32	0.30
	XGboost	Combine	-0.50	0.33	0.30
	XGboost	EN	-0.46	0.33	0.29
Task-3	Bert	Combine	-2.34	0.00	0.17
	Multilingual Bert model	ES	-2.29	0.00	0.17
	Multilingual Bert model	EN	-2.40	0.00	0.15

The table 4 include the Results given by the CLEF itself on test Dataset it include ICM-Hard which is a similarity function that generalizes Pointwise Mutual Information (PMI), and can be used to evaluate system outputs in classification problems by computing their similarity to the ground truth categories, ICM-hard Norm as well as F1-Score of Yes.

Table 5
Ranking achieved by runs of TASK 1, TASK 2, and TASK 3 on test dataset

Task	Model	Language	Rank
Task-1	Random Forest	Both	51
	Random Forest	ES	51
	Random Forest	EN	52
	XGboost	Both	45
	XGboost	ES	45
	XGboost	EN	47
Task-2	Random Forest	Both	35
	Random Forest	ES	35
	Random Forest	EN	32
	XGboost	Both	36
	XGboost	ES	36
	XGboost	EN	34
Task-3	Bert	Both	33
	Multilingual Bert	Es	33
	Multilingual Bert model	EN	31

The table 5 presents the performance rankings achieved by different models across three distinct tasks (Task-1, Task-2, and Task-3) as evaluated in the CLEF (Conference and Labs of the Evaluation Forum) runs. Each row specifies the type of model used (Random Forest, XGboost, Bert, Multilingual Bert) and the language dataset utilized (combined, Spanish, English). The ranks, ranging from 31 to 52, indicate how well each model performed relative to others within the same task and language context. Lower

ranks signify better performance, showcasing which models and language configurations excelled in the evaluation scenarios.

5. Conclusion

In conclusion, our study explored sexism classification using both traditional ensemble models (Random Forest and XGBoost) and a pretrained large language model (multilingual BERT). For the machine learning models, text encoding techniques such as TF-IDF and Bag of Words (BoW) were utilized. The multilingual BERT (mBERT) model outperformed all other models across the tasks, except for Task 2 where XGBoost demonstrated similar performance.

Despite the overall success of mBERT, the performance of all models was notably poorer in Task 2 and Task 3. This underperformance can be attributed to the highly imbalanced nature of the data, which posed significant challenges for model training and prediction accuracy. The imbalance likely resulted in a bias towards the majority class, reducing the models' ability to correctly classify the minority class instances. Nevertheless, this study provides valuable insights into the application of advanced machine learning techniques and pre-trained models for sexism classification in text.

Looking forward, research should expand comparisons among specialized large language models (LLMs) for English and Spanish to capitalize on language-specific nuances and improve classification accuracy. An effective strategy involves a language-specific approach where tweets are first categorized by language and then processed using dedicated pretrained models (e.g., English and Spanish). This method aims to enhance model performance by adapting analyses to each language's unique characteristics while maintaining overall classification integrity. While multilingual models like mBERT have performed well, future investigations could explore pretrained models exclusively trained on English and Spanish to mitigate challenges associated with multilingual training, such as nuanced language contexts and data biases. These specialized models are expected to enhance accuracy in sexism detection and other text classification tasks by better addressing specific linguistic nuances.

In summary, our study underscores the potential of advanced machine learning techniques and pretrained models in addressing sexism in social networks. Future research should prioritize refining model capabilities and addressing data challenges to enhance the effectiveness and applicability of sexism detection algorithms.

References

[1] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A multi-task and multilingual model for sexism identification in social networks, in: IberLEF@SEPLN, 2021.

[2] R. N. Montufar Mercado, Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques (2019).

[3] D. Eliezer, B. Major, W. B. Mendes, The costs of caring: Gender identification increases threat following exposure to sexism, Journal of Experimental Social Psychology 46 (2010) 159–165.

[4] S. Frenda, B. Ghanem, M. Montes-y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, Journal of intelligent & fuzzy systems 36 (2019) 4743–4752.

[5] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.

[6] R. Devadath, E. C. Alex, K. Sreeja, S. Abhishek, T. Anjali, A comparison of multinomial naive bayes and xg boost for sentiment analysis and bias detection in tweets, in: 2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), IEEE, 2024, pp. 1–7.

[7] B. Devi, V. G. Shankar, S. Srivastava, K. Nigam, L. Narang, Racist tweets-based sentiment analysis using individual and ensemble classifiers, in: Micro-Electronics and Telecommunication Engineering: Proceedings of 4th ICMETE 2020, Springer, 2021, pp. 555–567.

[8] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, arXiv preprint arXiv:2111.04551 (2021).

[9] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[10] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.