Automation of Profile Reporting System for Misogyny Identification

. **Somavarapu Jahnavi**
Computer Science & Engineering
VNRVJIET
Hyderabad, India

Marru Manogna
Computer Science & Engineering
VNRVJIET
Hyderabad, India

Sayini Anirudh
Computer Science & Engineering
VNRVJIET
Hyderabad, India

B. Santhosh Vara Siddu
Computer Science & Engineering
VNRVJIET
Hyderabad, India

Shaik Shoyeb Ahmed
Computer Science & Engineering
VNRVJIET
Hyderabad, India

***Abstract* — In most cases, women have been centered in critical situations unnecessarily in digital media. Responsible citizens can stop this by reporting such disparities in social media. On many social media platforms, based on the public voice (no. of reports), the higher authority will take action. This paper provides a summary of how we are preventing misogyny situations by implementing automation of the reporting process from the user end rather than expecting action from the higher authority after the damage has occurred. Our main job is to identify misogynist memes. Memes should be classified as misogynous or not, and misogyny should be divided into sorts such as stereotype, shaming, violence, and objectification.**

Keywords: quantizer, malicious, abnormalities, Bi-GRU, tinker, OpenAI, CLIP, corpus, paradigm, misogyny, Roberta, PerceiverIO, URLVoid, TrendMicro, linguistics, fusion.

I. INTRODUCTION

Online, women are prominent, especially on image-based platforms like Instagram and Twitter. The internet has opened up opportunities for women, but the same prejudice and discrimination that exist outside also exist online in the form of offensive material directed at them. Image macros, sometimes known as "memes," are a common communication technique on social networking sites. An internet meme is often a picture with superimposed text that was added later by the meme creator with the primary intention of being humorous and/or sarcastic.

While many memes are created only for laughs, some memes also possess pernicious objectives. Few people who are acquainted with the format would be startled to discover that memes may be used as a tool to spread violence and sexism online, furthering gender inequality and sexual stereotypes offline.

Monitoring and managing user profiles regularly is the key to reducing the circumstances of sexism in social media. The accounts of users who attempt to upload anti-women content frequently and with justification will be permanently deleted.

II. RELATED WORKS

Aditya Vailaya., [1] Under the restriction that the test the picture does belong to one of the classes, use binary Bayesian classifiers to try to extract high-level ideas from low-level visual attributes. Consider the hierarchical categorization of vacation photos: at the top level, photos are categorized as indoor or outdoor; outside, photos are further divided into city or landscape; and lastly, a portion of landscape photos are divided into classes for sunsets, forests, and mountains. It was shown that a compact vector quantizer might adequately express the class-conditional densities of the features needed by the Bayesian approach (Its preferred size is determined by modified MDL criteria).

Rima Masri., [2] In this research, a technique for automatically identifying fraudulent advertisements is proposed and put into practice. For the goal of detecting dangerous adverts, it uses three separate online malware domain detection systems (VirusTotal, URLVoid, and TrendMicro) and provides the number of advertisements found using each system.

Elena Shushkevich., [3] In order to identify sexism in messages taken from the Twitter platform, In this article, a method based on a combination of Naive Bayes, Support Vector Machines, and Logistic Regression models were presented.

Alessandra Teresa Cignarella., [4] The suggested framework has been tested on an Italian dataset based on Sentence Embeddings and Multi-Objective Bayesian Optimization. Here, they concentrated on the advantages and disadvantages of using pre-trained language as well as the role that Bayesian optimization plays in the issue of biased predictions.

Goenaga., [5] The recurrent neural network and convolutional neural network (CNN) are two widely used

models for these modelling tasks (RNN), which use quite different approaches to interpreting natural languages. In this study, they mainly on the RNN technique, which makes use of a Bidirectional Long Short Term Memory (Bi-LSTM) and Conditional Random Fields (CRF), and we assessed the suggested architecture on a task for identifying irregularities (text classification).

John Cardiff., [6] devised a method to identify sexism in tweets obtained from the Twitter website that incorporates Logistic Regression, Naive Bayes models, and Support Vector Machines.

Debbie Ging., [7] In order to find instances of sexism in the online slang lexicon Urban Dictionary, developed by the public, this research uses deep learning techniques. To identify sexist speech, the performance of two deep learning techniques(Bi-LSTM and Bi-GRU) has been evaluated, against that of more traditional machine learning techniques, such as logistic regression, Naive-Bayes classification, and Random Forest classification. They discovered that in contrast to the other strategies investigated, both deep learning algorithms are more accurate at spotting sexism in the Urban Dictionary.

Arjun Roy., [8] This task's goal was to foresee how online text posts or comments would propagate violence. The datasets were made available in two languages: Hindi and English. We provided one system for each of these languages. Individual models in both systems were created using ensembles of Convoluted Neural Networks (CNN) and Support Vector Machines (SVM).

Lei Chen., [9] utilized recently developed Transformer models that were pre-trained on large data sets (mainly by self-supervised learning) to provide extremely effective visual (V) and linguistic (L) characteristics. Specifically, we obtained coherent V and L features using the OpenAI CLIP model before making binary predictions with a logistic regression model. Second, by adhering to the data-centric AI approach, emphasis should be placed on data rather than model tinkering.

José Antonio García-Díaz., [10] On the one hand, misogynistic tweets on Twitter were found using applied sentiment analysis and social computing tools. On the other hand, created the Spanish MisoCorpus-2020, a well-proportional collection of written texts about misogyny in Spanish, and partitioned it into three divisions based on violence against women, common properties based on misogyny, as well as, pestering females through messages in Spanish and Latin America.

Niloofar Safi Samghabadi., [11] The task's data is supplied in three languages: Bengali, Hindi, and English. Data instances are categorized into aggressiveness classes such as Overtly Aggressive, Covertly Aggressive as well as Not Aggressive.On the other hand, categorized into two main misogyny classes: Non-Gendered and Gendered. Data for the work is provided in Bengali, Hindi, and English.

Endang Wahyu Pamungkas., [12] First, by developing a unique approach and carrying out a thorough assessment of this assignment, explore the key characteristics to spot misogyny and the problems that add to its difficulties. Secondly, carry out several cross-domain categorization studies to investigate the connection between sexism and other abusive language patterns. Finally, cross-lingual classification experiments were conducted to test the effectiveness of sexism detection in a bilingual environment.

Shardul Suryawanshi., [13] To determine if a specific meme is offensive or not, combine the two modalities. Used the memes associated with the U.S. presidential election held in 2016 to construct the MultiOFF multimodal meme dataset for rude content identification as there was no publically accessible dataset for such purposes. Using the MultiOFF dataset, a classifier was subsequently created for this purpose. To compare it to a baseline of only text and images, The visual and text modes were combined using an early fusion method.

Abdullah Y. Murad., [14] In this study, the method of identification of an Arabic word for sexism detection in Arabic tweets is given. The Arabic Levantine Twitter Dataset for Misogynistic is used to assess the suggested method, which achieved recognition accuracy for multi-class and binary tasks of 90.0% and 89.0%, respectively. The suggested method appears to help offer workable, intelligent ways for identifying Arabic misogyny on social media.

S. Rajeskannan., [15]Classification engine is displayed by an amalgamated model that is a combination of the feature extraction engine and a social media engine consisting of datasets using input raw texts. For CB identification, context, user comments, and psychological properties were extracted from the feature extraction engine. An artificial neural network (ANN) is used to classify the data, and the CB Identification may get rewards or penalties by an evaluation system where the classification engine has access to an evaluation system. Deep Reinforcement Learning (DRL), which boosts classification performance, is used for assessment.

Giuseppe Attanasio., [16] To solve tasks, utilize Perceiver IO to combine multimodal late streams with unimodal ones. Created unimodal embeddings using RoBERTa (text transcript) and Vision Transformer (picture). Additionally, face and demographic identification, picture captioning, adult material identification, and web entities were utilized to improve the depiction of the input. Perceiver IO is being used for the first time in this investigation to combine visual modalities and text.

Table 1: Table showing different methodologies, pros, cons, and the results obtained in this literature survey

S.No	Title	Methodology	Pros/Cons	Year
1	[1] Image Classification for Content-Based Indexing	Bayesian Framework, Vector Quantization for Density Estimation	PROS The accuracy rises if features are fixed since categorization is dependent on features. CONS The count of classes reduces as features grow, and they are based on both individual traits and combinations of features that are thought to be independent.	2015
2	[2] Automated Malicious Advertisement Detection using VirusTotal, URLVoid, and TrendMicro	Utilizing TrendMicro, VirusTotal, and URLVoid to Find Malvertisements	PROS After extracting URLs, the URLVoid has the greatest accuracy rate for detection. CONS The total proportion of the genuine positive is impacted by TrendMicro (malicious and the system classified it as malicious).	2017
3	[3] Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team	SVM, model ensembles, and carrying out a number of tasks such as pre-processing, model construction, and embedding the created models in one ensemble.	PROS Achieving a high degree of classification parameter estimate requires quick computations and little in the way of training data.. CONS As the number of classes is lowered, the model's efficiency as applied to the Misogyny Identification categorization drops.	2018
4	[4] Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon	Dialectal features from Linear and RBF kernel SVM, such as a multilingual hate dictionary, and structural features.	PROS The target's gender has been attained. It is determined whether males or women are participating.	2018

			CONS The disadvantages of RBF kernels are their high computational cost and worse performance in large and sparse feature matrices.	
5	[5] Automatic Misogyny Identification Using Neural Networks	Pre-trained word embeddings in the Bi-LSTM	PROS When it comes to feature selection, CRF is sufficiently versatile. CONS It won't function with CRF if the terms weren't known in the sample of training data.	2018
6	[6] Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team	Ensemble of logistic regression, SVM, and naive Bayes, with tf-id	PROS Highest accuracy, was achieved with the least amount of preprocessing labour. CONS A lack of understanding In other words, users have a hard time interpreting the knowledge.	2018
7	[7] A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary	Random Forest, Logistic Regression, and Naive Bayes Bi-GRU and Bi-LSTM	PROS The greatest outcomes are accuracy and sensitivity came from Bi-GRU and Bi-LSTM, whereas the finest outcomes of specificity came from Random Forest. CONS Without the Dl, the outcomes of the conventional ML approaches are quite poor.	2020
8	[8] An Ensemble approach for Aggression Identification in English and Hindi Text	To carry out the classification job, they used a neural architecture based on BERT with the word and distributional level embeddings.	PROS By using BERT Model, They processed a more enormous amount of text and language.	2020

			CONS NGEN classification is poor in three languages (Hindi, English, and Bengali)	
9	[9] Multimedia Misogyny Detection By Using Coherent Visual and Language Features from CLIP Model and Data-centric AI Principle.	Transformer models that have already been trained, such as the fine-tuning BERT model, the universal sentence encoding (USE) embedding, and the SBERT, and CLIP models	PROS In fact, performance is better when sentence-level representations and visuals are used. CONS Better results were obtained with an LR model alone than with more complex models.	2020
10	[10] Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings	Random forest, sequential minimal optimization (SMO), LSVM	PROS the ability to recognize sexism and effectively use binary classification CONS When dealing with multi-class classification issues, LSVM was unsuccessful.	2020
11	[11] Aggression and Misogyny Detection using BERT: A Multi-Task Approach	utilised many layers, including the Classification layer, Bert layer, Attention layer, and Fully-connected layer. the output of two distinct classification layers, one for identifying sexism and another for predicting aggressiveness class.	PROS The aggregate results demonstrate that sexism is easier to detect than antagonism in all available languages. CONS Across all the languages, the performance for CAG (Covertly Aggressive) is the lowest, indicating that it is the most difficult aggressiveness class to recognize.	2020
12	[12] Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study.	To fill the gap in Automatic Misogyny Identification in low-resource languages, a model was developed based on BERT and LSTM.	PROS Overall results show that the BERT-based model is the most successful model in the cross-lingual setting experiment. CONS The algorithm does not operate optimally when tested on data from AMI	2020

			and trained on data from other abusive events.	
13	[13] Multimodal Meme Dataset (Multi OFF) for Identifying Offensive Content in Image and Text	An early method of fusing text and images was used and its efficiency was tested by contrasting it with a baseline that used solely text and images.	PROS The only model that includes local embeddings is DNN. As a result, it outperformed other models by achieving better accuracy and an F1 score. CONS Should give textual characteristics more weight while integrating them with the meme's visual components.	2020
14	[14] AI-based Misogyny Detection from Arabic Levantine Twitter Tweets	With word and word embedding methods, the Arabic text is rendered. To identify sexism in Arabic, the most recent deep learning BERT approach is employed.	PROS Linear SVC model has achieved the highest accuracy among its peers. The results demonstrate the low performance of the Random Forest Classifier model. CONS The dataset lacked equilibrium. There are just 17 comments in the sexual harassment class, therefore there is very little opportunity to understand the pattern for these classes.	2021
15	[15] Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking	Artificial neural networks, feature selection, and information gain are all part of the DRL Algorithm for Reward-Penalty Decisions.	PROS The suggested ANN's accuracy has increased thanks to the DRL Algorithm. CONS The feature selection and model will not consider the latest CB trends.	2021
16	[16] Using Perceiver IO for Detecting Misogynous Memes with Text and Image Modalities	Perceiver IO was employed as a multimodal late fusion layer for multi-task learning, and they constructed a multimodal late fusion to jointly learn from several modalities.	PROS Adds semantic information to the meme, such as the image description, facial and demographic information, adult	2022

			content detection, and web entities. CONS The misogyny of target labels is balanced, while the other categories are unbalanced. Unbalance is a little more obvious than it was on the training set.	

III. CONCLUSION

According to the aforementioned literature review, several researchers have employed a variety of techniques to comprehend and automate the misogyny detection system. The integration of technology to make social media toxic-free is a theme that emerges in every study. Additionally, several techniques are described, including TFBertModel, TFViT, and ViTFeatureExtractor. Each study appears to concentrate on a different topic, such as toxicity identification, misogyny detection automation, meme classification, picture classification, etc. All of the techniques have demonstrated a respectable level of effectiveness when classifying memes. Better outcomes can be obtained by increasing financing and exercising control over a social media network.

IV. REFERENCES

[1] Vailaya, A., Figueiredo, M. A., Jain, A. K., & Zhang, H. J. (2001). Image classification for content-based indexing. IEEE transactions on image processing, 10(1), 117-130.

[2] Masri, R., & Aldwairi, M. (2017, April). Automated malicious advertisement detection using virustotal, urlvoid, and TrendMicro. In 2017 8th International Conference on Information and Communication Systems (ICICS) (pp. 336-341). IEEE.

[3] Cardiff, J., & Shushkevich, E. (2018). Misogyny detection and classification in English tweets: the experience of the ITT team. In Proc. EVALITA (p. 182).

[4] Endang, W. P., Alessandra, T. C., Valerio, B., & Viviana, P. (2018). Automatic identification of misogyny in English and Italian tweets at evalita 2018 with a multilingual hate lexicon. In CEUR Workshop Proceedings (Vol. 2263, No. 1, pp. 1-6). CEUR-WS.

[5] Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., Ilarraza, A.D., Ezeiza, N., Oronoz, M., Pérez, A., & Perez-de-Viñaspre, O. (2018). Automatic Misogyny Identification Using Neural Networks. IberEval@SEPLN.

[6] Caselli, T., Novielli, N., Patti, V., & Rosso, P. (2018, December). Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team. In Proceedings of the Final Workshop (Vol. 12, p. 13).

[7] Lynn, T., Endo, P. T., Rosati, P., Silva, I., Santos, G. L., & Ging, D. (2019, June). A comparison of machine learning approaches for detecting misogynistic speech in the urban dictionary. In 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA) (pp. 1-8). IEEE.

[8] Roy, A., Kapil, P., Basak, K., & Ekbal, A. (2018, August). An ensemble approach for aggression identification in English and Hindi text. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018) (pp. 66-73).

[9] Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., ... & Sorensen, J. (2022, July). SemEval-2022 Task 5: Multimedia automatic misogyny identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 533-549).

[10] García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. Future Generation Computer Systems, 114, 506-518.

[11] Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., & Solorio, T. (2020, May). Aggression and misogyny detection using BERT: A

multi-task approach. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 126-131).

[12] Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: a multilingual and cross-domain study. Information Processing & Management, 57(6), 102360.

[13] Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020, May). Multimodal meme dataset (MultiOFF) for identifying offensive content in images and text. In Proceedings of the second workshop on trolling, aggression and cyberbullying (pp. 32-41).

[14] Muaad, A. Y., Davanagere, H. J., Al-antari, M. A., Benifa, J. B., & Chola, C. (2021, September). AI-based misogyny detection from Arabic Levantine Twitter tweets. In Computer Sciences & Mathematics Forum (Vol. 2, No. 1, p. 15). MDPI.

[15] Yuvaraj, N., Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S. M. G. S. M. A., ... & Masud, M. (2021). Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. Mathematical Problems in Engineering, 2021.

[16] Attanasio, G., Nozza, D., & Bianchi, F. (2022, July). MilaNLP at semeval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 654-662).iction error estimation," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 3, pp. 569-575,2010.