MAMITA: Benchmarking Misogyny in Italian Memes

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi and Aurora Saibene

University of Milano-Bicocca, Milan, Italy

Abstract

This paper introduces **MAMITA**, a novel Italian multimodal benchmark dataset developed for the automatic detection of misogynistic content in online media, with a specific focus on memes. The dataset comprises 1880 memes sourced from popular social platforms—Facebook, Twitter, Instagram, Reddit—and meme-centric websites, selected using misogyny-related keywords covering a wide range of manifestations including body shaming, stereotyping, objectification, and violence. A key feature of this benchmark is its dual annotation strategy: all memes were independently labeled by both domain experts and a pool of 232 crowd annotators. This approach resulted in two parallel sets of annotations that reflect differing labeling perspectives. For each meme, labels include a binary classification (misogynistic or not), the type of misogyny, and its intensity. Beyond categorical labels, the dataset incorporates perspectivist metadata, capturing individual annotators' perceptions of misogyny along with their demographic and socio-cultural background, including age, level of education, and social status. Each meme's textual content was also automatically transcribed to enable multimodal analysis. This enriched benchmark enables nuanced research on the automatic detection of misogynistic content in online social media and supports investigations into how perceived misogyny varies across annotator profiles, allowing us to address the urgent challenge related to the diffusion of hateful content against women.

Warning: this paper includes examples that may be offensive or harmful.

Keywords

Misogynous Memes, Italian Benchmark, Expert vs Crowd Annotation, Perspectivism

1. Introduction

In recent years, the proliferation of user-generated content on social media has intensified the creation of hateful content against women not only using textual messages that can implicitly or explicitly contain harmful content, but also from a multimodal perspective[1]. Among the diverse forms of online expression, memes have emerged as viral communication tools, which can subtly convey harmful ideologies thanks to their combination of visual and textual elements. This kind of digital violence can be an extension or a precursor to physical violence, stalking and harassment, but it can also be a way to punish, abuse or silence women, increasing the isolation of victims (Council of Europe, 2021) [2]. Through the combination of apparently innocuous images coupled with harmless superimposed text, misogynous memes can be easily created and spread, normalizing and trivializing detrimental stereotypes, objectification, and marginalization of women. Their viral nature, usually due to the ironic message behind, contributes to their rapid spread across several media platforms, also fueling those communities that reinforce misogynistic ideologies.

Despite growing societal awareness and policy efforts aimed at addressing such an issue, the automatic detection of multimodal misogynistic content remains a significant challenge. A major limitation in the development of robust misogyny detection systems is the scarcity of high-quality, multimodal datasets that reflect the nuanced and subjective nature of such content. Misogyny can manifest in explicit or implicit forms, often relying on cultural references, irony, or layered symbolism.

The identification of this kind of abusive content is of paramount importance not only for protecting women and guaranteeing safe online environments, but also for eventually generating counter-narratives [2].

In this paper, we provide three main contributions:

1. **MAMITA** (**M**ultimedia **A**utomatic **M**isogyny **I**dentification in i**TA**lian), a novel Italian benchmark focused on misogynistic content in memes, which covers diverse forms of gender-based hate such as body shaming, objectification, stereotype, and violence.
2. **Dual annotation strategy** involving both domain experts and crowd annotators, enabling comparative analysis of labeling perspectives and improving the robustness of misogyny detection.
3. **Perspectivist annotation**, capturing for each annotator perceived misogyny along with demographic and socio-cultural background such as age, education, and social status, to support re-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]

elisabetta.fersini@unimib.it (E. Fersini); elisabetta.fersini@unimib.it (F. Gasparini); giulia.rizzi@unimib.it (G. Rizzi); aurora.saibene@unimib.it (A. Saibene)

[1] https://www.unwomen.org/sites/default/files/2024-10/a-79-500-sg-report-ending-violence-against-women-and-girls-2024-infographic-and-recommendations-en.pdf

[2] https://rm.coe.int/study-on-effectiveness-risks-and-potentials-of-using-counter-and-alter/1680b40775

(a) Shaming (b) Stereotype (c) Objectification (d) Violence

Figure 1: Examples of misogynous memes.

search on disagreement in hate speech perception and detection.

The paper is organized as follows. In Section 2, related works are presented. In Section 3, the proposed benchmark is described, detailing the two types of annotations, i.e., experts and crowd. In Section 4, insights from human and multimodal models are reported. Finally, in Section 5, conclusions are outlined.

2. Related Work

The automatic detection of hate speech, and misogyny in particular, has received growing attention in Natural Language Processing (NLP). Early efforts have primarily focused on text-based misogyny detection [3], using datasets sourced from Twitter and Reddit. For instance, regarding the multilingual settings, several benchmark datasets have been proposed in the literature to cover multiple languages. A few representative benchmarks are denoted by HATEVAL [4] focused on English and Spanish, BAJER [5] for the Danish language, BIASLY [6] focused on movie subtitles and colloquial expressions in North American film, ArMIS [7] for the Arabic language, and EXIST [8, 9] for dealing with English and Spanish sexist expressions.

Regarding the Italian language, we can summarize two main benchmarking text-related initiatives, i.e., AMI [10, 11, 12] and PejorativITy [13]. AMI (Automatic Misogyny Identification) represents a set of benchmark datasets that, starting from the initial challenge at Evalita 2018, have led to three main annotated corpora, i.e., AMI@Evalita 2018, AMI@Evalita 2020, and AMI-PRF. The AMI@Evalita 2018 dataset introduced in [10] provided one of the first benchmarks for detecting misogynistic language on social media in English and Italian tweets. Its extension presented at the AMI@Evalita 2020 [11] denotes an extension of the former benchmark to also capture aggressiveness. Lastly, AMI-PRF [12] is the most recent dataset of tweets annotated for both misogyny and professional categories. A further contribution is represented by PejorativITy [13], an Italian tweet corpus annotated at word level for pejorativity, and at the sentence level for misogyny.

While these efforts advanced text-based detection, they did not address the complexity of multimodal content such as memes, which often rely on implicit visual cues, humor, and cultural references to communicate harmful messages. Among the general hateful meme benchmarks, we can highlight four main initiatives focused on the English language, i.e., Facebook Hateful Memes [14], Memotion2 [15], Harmful Memes[16], MultiOFF [17], and Intervening Cyberbullying in Multimodal Memes (ICMM) [18]. However, these benchmarks do not capture the specificity of misogyny, which often relies on gender norms, implicit bias, and culturally coded references that differ significantly from general offensive content or other forms of targeted hate (e.g., against immigrants or people with disabilities). Only a few benchmarks have been proposed to deal with the peculiarity of hate against women in a multimodal settings, i.e., MAMI [19] for the English language, MIMIC [20] for Hindi, EXIST [21, 22] for English and Spanish, and Dravidian corpus [23] focused on the Tamil and Malayalam languages.

Although all the previous initiatives represents a fundamental step towards the identification of hateful meme against women, to the best of our knowledge no benchmark dataset has been developed to specifically address misogynistic content in the Italian language, resulting in a remarkable gap in the resources available for the systematic investigation of this phenomenon within the Italian contexts. To this purpose, we propose **MAMITA** (**M**ultimedia **A**utomatic **M**isogyny **I**dentification in i**TA**lian), a novel benchmark dataset for the Italian language that focuses on misogynous memes,

composed of a wide range of multimodal expressions denoting body shaming, objectification, stereotyping, and violence. The dataset is developed using a dual annotation strategy that combines input from both domain experts and crowd annotators, enabling robust analysis of labeling perspectives.

3. MAMITA

The meme collection was primarily carried out using visual search engines such as Google Images and Pinterest, based on the keywords reported in Table 1. All the keywords have been defined to try to capture four main categories related to misogynous contents, i.e., body shaming, objectification, stereotyping, and violence. The websites considered are typically dedicated to meme sharing (e.g., *me.me* and *memedroid.com*), as well as Instagram accounts focused on themes related to femininity (e.g., *alpha woman* and *scaricatricidiporto*). Additional content was sourced from Facebook groups intentionally created for the dissemination of misogynistic memes (e.g., *facciaabuco*, *ignoranza soffocotti pecorina*, and *Io sono vaginatariano*). The initial dataset consisted of approximately 2,000 memes. Pornographic content, low-quality images, and items that could not be clearly categorized as memes were subsequently removed. Memes were also normalized to a maximum resolution of 640×640 pixels, preserving their aspect ratio. The final dataset comprises 1880 memes, with the textual content transcribed using Optical Character Recognition (OCR) tools (https://www.onlineocr.net/). Examples of misogynous memes available in the MAMITA dataset are reported in Figure 1. The dataset has been subsequently labeled by two distinct groups, i.e., expert and crowd annotators. The full dataset can be accessed by filling in the form https://forms.gle/5Xz1gcxJdrh6GHnq5.

3.1. Expert Annotation

For what regards the annotation process performed by the **experts**, we involved two male and three female annotators. In order to label each meme, they adopted the definitions originally provided in [19], opportunely adapted for covering the multimodal scenario. Each meme was reviewed by one male and two female experts. Each expert involved in the evaluation process analyzed the memes, classifying them as either misogynistic or non-misogynistic. In cases where a meme was perceived as misogynistic, evaluators were also asked to specify the type of misogyny, selecting among violence, body shaming, stereotyping, and objectification. In cases of uncertainty about the categorization, evaluators were allowed to select multiple types of misogyny.

The annotation process performed by the experts has led to a full agreement in 81.43% of the memes, where in 70.86% of such cases the memes were labeled by the three annotators as misogynous. We computed Fleiss' Kappa statistics [24] to assess the level of agreement among the experts. The resulting score was 0.749, indicating a substantial inter-annotator reliability in the perception of memes. This value suggests a strong consistency in the evaluators' judgments, particularly in distinguishing between misogynistic and non-misogynistic content.

The annotations given by the experts have also been aggregated following a majority voting strategy to assign a final golden label about misogyny. The dataset labeled by the experts finally consists of 57.71% of misogynous and 42.29% of not misogynous memes. Regarding the category of misogyny, since multiple overlapping annotations were possible, the final dataset evaluated by the expert contains - among those memes considered as misogynous by the majority of the experts - 76.12% of the memes labeled as Objectification, 48.29% as Stereotype, 20.18% as Violence and 8.84% as Body Shaming by at least one annotator. Considering that multiple labels are allowed for the type of misogyny, the dataset is provided with soft labels denoting a probability distribution for each category.

3.2. Crowd Annotation

For what concerns the annotation process performed by the **crowd**, we prepared a proper Google Form and we engaged trusted voluntary annotators (from 4 to 10 labelers for each meme). The total number of volunteers involved is 231 (116 male, 110 female, and 5 non-responders). The most frequent age is between 25-34 years old, i.e., about 41% of the annotators. The native language is Italian for the 99% of the participants, while the remaining three annotators speak Italian fluently. The dataset was divided into groups of 40 memes each, balanced in terms of classification (20 misogynistic, 20 non-misogynistic) according to the experts' preliminary evaluations, to be subsequently evaluated by the engaged crowd annotators. The choice of presenting a limited number of memes is due to the fact that sensory habituation cause people to reduce their response to repeated or continuous stimuli over time [25].

Each meme was independently reviewed by a varying number of labelers. Each annotator labeled the memes as either misogynistic or non-misogynistic and, when applicable, selected the primary *Category* of misogyny that they perceived most together with the *Intensity* of figured out misogyny. Moreover, in order to provide a benchmark that is characterized by perspectivist information, we acquired a few variables to characterize the annotators. In particular, participants were required to provide a few information about themselves. Specifically, the following specific details have been required:

bitch (stronza)	fat (grassa)	milf (milf)
blondes (bionde)	female (femmina)	misogynist (misogino)
call girl (escort)	feminism (femminismo)	misogyny (misoginia)
cheap (squallida)	feminist (femminista)	nazifeminist (nazifemminista)
cheat (tradire / imbrogliare)	fuck (fottiti / scopare)	pregnancy (gravidanza)
clean (pulire)	girl (ragazza)	promiscuous (promiscua)
cleaning (pulizia)	girlfriend (fidanzata)	prostitute (prostituta)
cold (fredda)	girl power (potere femminile)	rape (stupro)
complicated (complicata)	girls (ragazze)	sandwich (panino)
cooking (cucinare)	gold digger (arrampicatrice sociale)	sex (sesso)
cougar (cugar)	harsch (dura / severa)	sexism (sessismo)
couple (coppia)	hooker (prostituta)	sexist (sessista)
crazy (pazza)	hore (puttana)	slut (zoccola)
cunt (cagna)	house (casa)	stupid (stupida)
dirty (sporca)	housewife (casalinga)	tits (tette)
dishwasher (lavastoviglie)	inferior (inferiore)	trixie (ragazza superficiale)
driving (guida)	kitchen (cucina)	unstable (instabile)
dumb (stupida)	lazy (pigra)	wife (moglie)
equal rights (pari diritti)	marriage (matrimonio)	witch (strega)
escort (escort)	Mars & Venus (Marte e Venere)	woman (donna)

Table 1
List of keywords used to collect the MAMITA benchmark dataset.

(1) Socio-Demographic Characteristics:

- **Gender**: male, female, not specified
- **Age**: 18-24, 25-34, 35-44, 45-54, 55-64, more than 65 years old
- **Nationality**: legal status of a person based on their country of citizenship
- **Native language**: language connection to family and cultural identity
- **Education level**: Primary school, Middle school, High school, Bachelor's degree, Master's degree, Postgraduate Specialization, or PhD.
- **Employment Status**: Student, Working Student, Worker, Unemployed, Retired, or Other.

(2) Individual Beliefs:

- **Subjective Social Status (SSS)**: we introduced a variable that has the goal to measure an individual's perception of his/her social position compared to others. To this purpose, we adopted the MacArthur scale introduced in [26]. Participants are asked to place themselves on a graduated scale consisting of ten steps, ranging from the highest to the lowest socioeconomic status. At the top of the scale (10) are individuals with the highest levels of income, education, and occupational prestige. At the bottom of the scale (1) are those with the lowest income, minimal education, and the least respected jobs, or who may be unemployed. This self-placement invites participants to express a subjective evaluation of their social position with respect to other members of the society
- **Political Orientation**: participants were invited to express their political orientation on a 7-point Likert scale, where 1 indicates *Far Left* and 7 *Far Right*
- **Religious Orientation**: Catholic, Protestant, Orthodox, Muslim, Jewish, Hindu, Buddhist, Atheist, Agnostic, Other
- **Sensitivity towards misogyny**: participants were invited to express their sensitivity towards misogynous content using a 7-point Likert scale, where 1 denotes *Not at all sensitive* and 7 *Extremely sensitive*.

(3) Meme awareness:

- **Familiarity with memes**: Yes/No response to whether they know what memes are
- **Frequency of meme visualization**: how often the participant encounters memes, using a 7-point Likert scale ranging from *Never* to *Very Often*
- **Primary source of meme stimuli**: social media, messaging apps, websites and forums, other.

Since the number of annotators varies for each meme, they have been finally labeled as misogynous if at least 50% of the annotators provided the misogynous label. Based on the crowd annotations, the resulting dataset consists of 58.82% misogynous and 41.17% non-misogynous memes. The annotation process led to full

agreement for 43.14% of the memes. If we focus on each class, 37.97% of the misogynous memes and 50.45% of the not misogynous ones show a full agreement, denoting (as expected) a higher disagreement on misogynous content. To evaluate the overall level of agreement, we also computed Krippendorff's Alpha statistic [27], which yielded a score of 0.43. While the percentage of full agreement suggests some level of consistency, the Krippendorff's Alpha value indicates that a substantial portion of the agreement may be attributable to chance, highlighting extremely subjective interpretation of what can be considered as misogynous. As for the specific categories of misogyny, the dataset includes 70.97% of misogynous memes labeled as objectification, 55.87% as stereotype, 30.47% as violence, and 22.47% as body shaming by at least one annotator. Also in this case the dataset is provided with soft labels denoting a probability distribution for each category derived through the crowd annotation process.

4. Insights from MAMITA

In this section, we present a twofold analysis of the MAMITA dataset. First, we investigate how socio-demographic and cognitive characteristics of human annotators—such as gender, age, and Subjective Social Status—influence the perception and labeling of misogynistic content. Then, we evaluate the performance of multimodal baseline models, specifically mCLIP and mBLIP, in detecting misogyny and disagreement in memes, providing a comparative perspective between human subjectivity and machine predictions.

4.1. Human Perspectives

To better understand how individual differences influence the perception of misogynistic content, we formulated three research questions.

[Q1] Does the perceived intensity of misogyny significantly differ between male and female annotators? The aim is to determine whether the observed differences in the perception of misogyny intensity between men and women are statistically significant or could be due to chance. To this purpose, the Welch t-test has been adopted, which does not assume the same variance between the two populations. In this specific case, the null hypothesis is that the two means of the perceived intensity are equal and that any observed difference in the data can be attributed to random error or natural sample variation, rather than to a real effect.

The Welch t-test is -13.98, where the negative sign indicates the direction of the difference since the mean of women is higher than that of men (5.07 vs. 4.29) and the absolute value indicates how large the difference is in terms of standard deviation, i.e., the larger the absolute value, the more statistically significant the difference. In this case, the p-value, which indicates the likelihood that this difference occurred by chance, is extremely low (2.14×10^{-43}). The results show a ***highly significant difference in the perception of intensity between men and women***, suggesting that the probability of observing such a difference by chance is asymptotic to zero.

[Q2] Do statistically significant differences exist among age groups to identify misogynistic content? The core idea is to assess whether the probability of judging content as misogynistic depends on the annotator's age group. For this purpose, we estimated both a Chi-Squared statistic and a Binary Logistic Regression, which verifies if there exists a relationship and estimates how much each age group affects the likelihood of judging content as misogynistic, respectively.

In our case, the p-value equal to 7.10 related to the Chi-Squared test denotes ***a statistically significant relationship between age and the misogyny judgment***. As an additional observation, we report in Table 2 the results of the Binary Logistic Regression where the dependent variable (misogynous or not) is binary.

Age	p-value	Odds Ratio
25-34	0.000	1.24
35-44	0.001	1.28
45-54	0.000	1.52
55-64	0.000	1.43
$\geq$65	0.002	1.50

Table 2
Results o the Binary Logistic Regression.

The independent variables are age categories, compared with a reference category 18-24 age group. We can easily note that the socio-demographic attribute related to the Age is significantly associated with the likelihood of labeling content as misogynistic, where all age groups compared to the baseline (18-24) are statistically significant (p-value < 0.01). Moreover, the Odds Ratios increase with age, particularly from age 45 and up. This indicates an increased probability of labeling content as misogynous as age increases (compared to the 18-24).

[Q3] Has the Subjective Social Status a significant relationship with the intensity of the perceived misogyny? To explore the relationship between individuals' perceived social standing and their sensitivity to misogynistic content, we computed the Spearman correlation between SSS and the perceived intensity of misogyny. In particular, for each annotator, we considered their self-reported SSS score obtained from the back-

ground questionnaire and calculated the average intensity of misogyny they assigned across all memes they annotated as misogynistic. This approach allowed us to assess whether annotators with differing self-reported social positions systematically varied in how strongly they perceived misogynistic content. Spearman's rank correlation was chosen due to its suitability for capturing monotonic relationships without assuming normality in the data distributions.

The Spearman correlation analysis between the Social Sensitivity Score and the perceived intensity of misogynistic content yielded a statistically significant positive correlation ($\rho = 0.209$, $p = 0.0015$). While the correlation is relatively weak, it indicates that annotators with a ***higher Social Sensitivity Score are slightly more likely to assign higher intensity of perceived misogyny***. This finding highlights the influence of annotator-level socio-cognitive traits on subjective annotation tasks and suggests the importance of modeling annotator variability when addressing harmful or sensitive content.

4.2. Multimodal Baseline Models

To assess the effectiveness of multimodal models in identifying misogynistic content and disagreement between annotators, we fine-tune two state-of-the-art architectures: mCLIP [3][28, 29] and mBLIP[4] [30]. These models leverage both visual and textual information from memes, enabling a comprehensive understanding of their content. Both the vision encoder and text decoder are trained jointly with a classification head, allowing the models to tailor their multimodal representations to the specific task of misogyny and disagreement detection on the MAMITA dataset. To provide a simple and consistent baseline for evaluation, we fine-tune both models by adding a linear classification layer on top of their original representations, without further architectural modifications[5]. The classifier is trained using binary cross-entropy loss and the Adam optimizer. To compare the baseline models, we measure Precision (P), Recall (R), and F-Measure (F1), distinguishing between the misogynous label (+) and the non-misogynous one (-) as well as the agreement label (+) vs the disagreement one (-). We adopt a 10-fold cross-validation approach to ensure robustness and generalizability of the evaluation.

The results reported in Table 3 highlight the performance of mCLIP and mBLIP in predicting misogynistic content, evaluated against both Crowd and Expert annotations. Overall, both models show good classification capabilities, with mBLIP consistently outperforming mCLIP across all metrics. In the Crowd setting, mBLIP achieves a higher average F1 score (0.79 vs. 0.70), demonstrating better balance between precision and recall for both misogynous and not misogynous labels. It is interesting to note that mBLIP's $F1^+$ (0.83) and $F1^-$ (0.76) suggest a strong ability to correctly identify both misogynistic and non-misogynistic content according to crowd judgments. Performance improves further when considering the Expert annotations. Both models exhibit higher F1 scores compared to the Crowd setting, with mBLIP again leading (Avg. F1 = 0.83 vs. 0.73 for mCLIP). This may indicate better alignment between the models' predictions and the expert labeling criteria, possibly due to more consistent or less ambiguous expert judgments. In both evaluation contexts, mBLIP proves to be the more robust of the two models, offering more reliable and accurate misogyny detection. These results suggest that state-of-the-art multimodal models, particularly mBLIP, can effectively capture harmful content signals when fine-tuned appropriately.

			Crowd				
Approach	P^+	R^+	$F1^+$	P^-	R^-	$F1^-$	$Avg.F1$
$mCLIP$	0.84	0.61	0.71	0.60	0.83	0.69	0.70
$mBLIP$	0.84	0.81	0.83	0.74	0.78	0.76	0.79
			Expert				
Approach	P^+	R^+	$F1^+$	P^-	R^-	$F1^-$	$Avg.F1$
$mCLIP$	0.86	0.63	0.73	0.63	0.86	0.73	0.73
$mBLIP$	0.88	0.82	0.85	0.77	0.85	0.81	0.83

Table 3
Misogyny prediction performance on Crowd and Expert labels.

			Crowd				
Approach	P^+	R^+	$F1^+$	P^-	R^-	$F1^-$	$Avg.F1$
$mCLIP$	0.00	0.00	0.00	0.57	1.00	0.72	0.36
$mBLIP$	0.00	0.00	0.00	0.57	1.00	0.72	0.36
$mCLIP^{(*)}$	0.44	0.52	0.48	0.58	0.49	0.53	0.50
$mBLIP^{(*)}$	0.42	0.69	0.53	0.55	0.28	0.37	0.45
			Expert				
Approach	P^+	R^+	$F1^+$	P^-	R^-	$F1^-$	$Avg.F1$
$mCLIP$	0.81	1.00	0.90	0.00	0.00	0.00	0.45
$mBLIP$	0.81	1.00	0.90	0.00	0.00	0.00	0.45
$mCLIP^{(*)}$	0.82	0.37	0.51	0.19	0.66	0.30	0.40
$mBLIP^{(*)}$	0.83	0.34	0.49	0.19	0.68	0.30	0.39

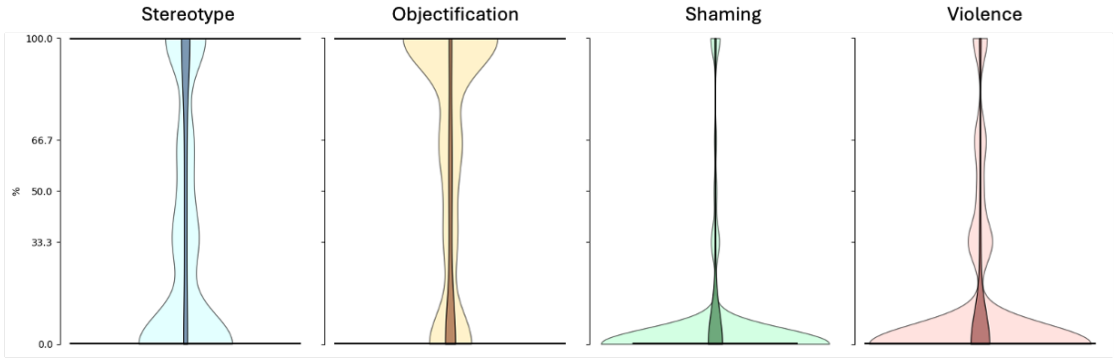
Table 4
Disagreement prediction performance on Crowd and Expert labels. $^{(*)}$ denotes models calibrated using the the Youden's J statistic.

Table 4 reports the performance of the considered baseline models in predicting disagreement between crowd and expert judgments, under two conditions: raw model outputs and outputs calibrated using the Youden's J statistic [31] to determine the best classification threshold on the probability distribution. When evaluating against the

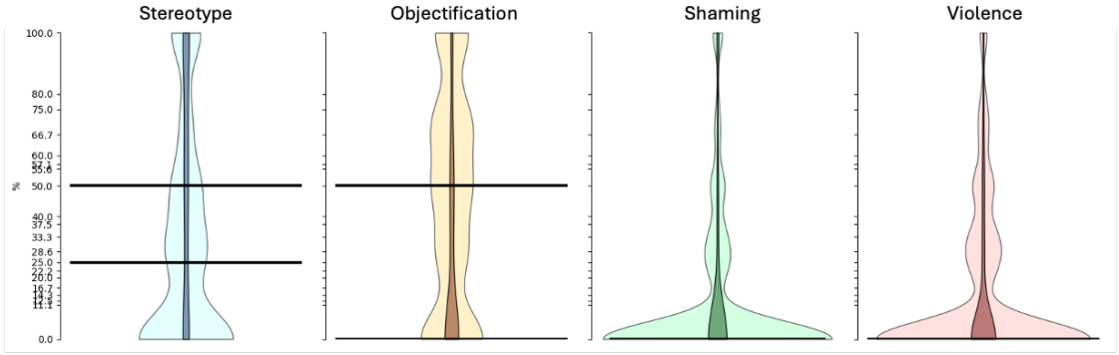
[3] https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

[4] https://huggingface.co/Gregor/mblip-mt0-xl

[5] To ensure reproducibility of our results, we report the main training parameters used: batch size = 4, classification threshold = 0.5, and number of training epochs = 5.



(a) Expert



(b) Crowd

Figure 2: Violin plots showing the distribution of misogynous data (bright colors) and corresponding prediction errors (dark overlays) for each label: Stereotype, Objectification, Shaming, and Violence. The plots illustrate the variability within the dataset and highlight the concentration and spread of errors relative to each label.

Crowd labels, both mCLIP and mBLIP perform poorly, assigning all instances to the negative class. However, applying the Youden correction significantly improves performance, increasing the average F1 from 0.35 to 0.50 for mCLIP and 0.45 for mBLIP. In the Expert setting, uncalibrated models exhibit an inverse pattern: perfect recall and high precision for positive labels (F1 = 0.90), but do not detect negative samples, again reflecting a strong prediction bias. The use of the Youden's threshold reduces such a bias ($F1^{-} = 0.30$), at the cost of reduced precision and recall on the positive class. Overall, these results highlight a key challenge in using pretrained multimodal models for subtle content moderation tasks: while default thresholds may lead to heavily skewed predictions, simple calibration strategies can significantly rebalance model behavior, though not without trade-offs.

We further analyzed models' errors to better evaluate models' performances, particularly considering the instances that were mislabeled by both classification models. A first analysis focuses on the evaluation of errors in misogyny identification with respect to the different types of misogyny. Figure 2 reports four violin plots corresponding to different misogyny categories, distinguishing between Experts and Crowd annotations. Each plot displays the distribution of a specific variable as a percentage[6] on the y-axis. The bright-colored regions represent the distributions within the whole dataset, while the darker-colored regions overlaid within each violin illustrate the distribution of the errors for each label. From the visual comparison, we can easily notice that:

- Stereotype and Objectification labels exhibit relatively symmetrical and balanced distributions with a moderate spread, indicating consistent distribution across a broad range of values. The error distributions for these labels are also centered, suggesting relatively low and uniform prediction errors.
- Shaming and Violence have a sharp, narrow dataset and error distributions, denoting a lot

[6]The percentage value has been computed with respect to the subset of data labeled as misogynous by the majority of annotators.

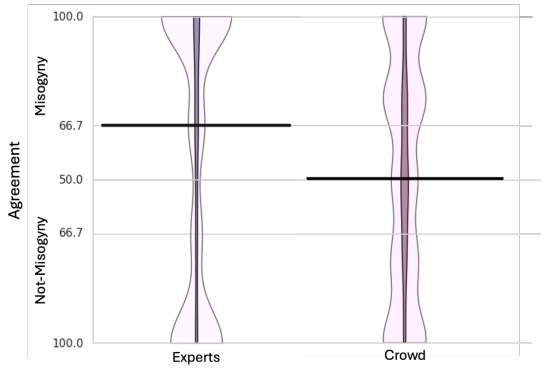


Figure 3: Violin plots showing the distribution of annotator agreement (y-axis, percentage) distinguishing between class label (Misogyny vs. Not-Misogyny) and annotation source (Experts vs. Crowd). The lighter area in each violin represents the full dataset distribution, while the darker overlay indicates the distribution of model prediction errors.

of misogynous memes not belonging to those classes.

By analyzing the shapes of the violin plots, we can notice that the violins dedicated to Shaming and Violence assume a shape broader at the basis, denoting a significant portion of misogynous memes not labeled with those types. Considering all the misogyny types, we can notice that the Expert plot is consistent in shape with the Crowd one for all the types, indicating a general ability of the crowd annotators in recognizing all the misogyny types.

Subsequently, we evaluated models' ability in detecting misogynistic content with respect to disagreement between annotators. Figure 3 reports two violin plots of the agreement among annotators along with the prediction error distributions for misogyny classification, distinguishing between Expert and Crowd annotators. The y-axis represents annotator agreement as a percentage, with higher values indicating stronger consensus among annotators, both on the Misogynous and Non-Misogynous labels. Each violin, representing the Expert and the crowd evaluation respectively, is divided into two layers: the lighter area represents the distribution of the full dataset, while the darker overlay highlights the distribution of the model's prediction errors. It is easy to notice that the Expert-dedicated violin assumes an hourglass shape, denoting a tendency for Experts to agree on both classes. The crowd plot instead shows a more uniform distribution, denoting a greater variability in the disagreement between crowd annotators. In both cases, the error distribution appears to be consistent and unrelated to the disagreement distribution. These patterns indicate that model errors are not influenced by the degree of annotator agreement. As part of future work, we plan to conduct a more in-depth qualitative error analysis, with a specific focus on identifying the most challenging archetypes of controversial or ambiguous memes, following the approach proposed in [32], to better understand the limitations of current models and highlight open challenges in the detection of misogyny in Italian.

5. Conclusions

In this paper, we presented a novel Italian multimodal benchmark dataset designed to support the automatic detection of misogynistic memes in online social media. The dataset emphasizes diversity in content and labeling perspectives, offering a comprehensive view of how misogyny is manifested and perceived across different annotator groups. The proposed benchmark, collected using a variety of popular platforms and focusing on a wide spectrum of misogynistic expressions, ensures a broad coverage of the phenomenon. Moreover, the dual annotation strategy, which includes both domain experts and crowd annotators, provides an opportunity to investigate the discrepancies in perceiving contents, therefore improving the robustness of future automatic detection systems that account for perspectivism.

Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. This work has also been supported by ReGAInS, Department of Excellence. The authors would also like to thank the significant contributions of the master's students Annalisa Bachir, Gökalp Recep Boz, Gaia Campisi, Marco Cervelli, Lisa Cocchia, Francesca Frigerio, Rosa Gotti, Monica Mantovani, Matteo Parisi, Emma Salvadori, whose dedicated efforts were fundamental to the development and compilation of the MAMITA dataset.

References

[1] C. Bosco, E. Ježek, M. Polignano, M. Sanguinetti, Preface to the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025, pp. –.

[2] The Council of Europe, 6th general report on grevio's activities: Group of experts on action against violence against women and domestic violence,

2024. URL: https://rm.coe.int/6th-general-report-on-grevio-s-activities/1680b5cbe8.

[3] S. Hewitt, T. Tiropanis, C. Bokhove, The problem of identifying misogynist language on twitter (and other online social spaces), in: Proceedings of the 8th ACM Conference on Web Science, 2016, pp. 333–335.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63.

[5] P. Zeinert, N. Inie, L. Derczynski, Annotating online misogyny, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3181–3197.

[6] B. Sheppard, A. Richter, A. Cohen, E. Smith, T. Kneese, C. Pelletier, I. Baldini, Y. Dong, Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 427–452.

[7] D. Almanea, M. Poesio, ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2282–2291.

[8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[9] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: European Conference on Information Retrieval, Springer, 2023, pp. 593–599.

[10] E. Fersini, D. Nozza, P. Rosso, et al., Overview of the evalita 2018 task on automatic misogyny identification (ami), in: CEUR workshop proceedings, volume 2263, CEUR-WS, 2018, pp. 1–9.

[11] E. Fersini, D. Nozza, P. Rosso, et al., Ami@ evalita2020: Automatic misogyny identification, in: Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), 2020.

[12] A. Cascione, A. Cerulli, M. M. Manerba, L. Passaro, Women's professions and targeted misogyny online, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 182–189.

[13] A. Muti, F. Ruggeri, C. Toraman, A. Barrón-Cedeño, S. Algherini, L. Musetti, S. Ronchi, G. Saretto, C. Zapparoli, PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 12700–12711.

[14] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624.

[15] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. DaS, T. Chakraborty, A. Sheth, A. Ekbal, et al., Memotion 2: Dataset on sentiment and emotion analysis of memes, in: Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR, 2022.

[16] S. Sharma, M. S. Akhtar, P. Nakov, T. Chakraborty, DISARM: Detecting the victims targeted by harmful memes, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1572–1588.

[17] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 32–41.

[18] P. Jha, R. Jain, K. Mandal, A. Chadha, S. Saha, P. Bhattacharyya, MemeGuard: An LLM and VLM-based framework for advancing content moderation via meme intervention, in: L.-W. Ku, A. Martins,

V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 8084–8104.

[19] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.

[20] A. Singh, D. Sharma, V. K. Singh, Mimic: misogyny identification in multimodal internet content in hindi-english code-mixed language, ACM Transactions on Asian and Low-Resource Language Information Processing (2024).

[21] 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 tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.

[22] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: European Conference on Information Retrieval, Springer, 2025, pp. 442–449.

[23] B. R. Chakravarthi, R. Ponnusamy, S. Rajiakodi, S. P. M. Chinnan, P. Buitelaar, B. Sivagnanam, A. KA, Findings of the shared task on misogyny meme detection: Dravidianlangtech@ naacl 2025, in: Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, 2025, pp. 721–731.

[24] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.

[25] V. Tarantino, N. Passerello, A. Ben-Sasson, T. Y. Podoly, A. Santostefano, M. Oliveri, L. Mandolesi, P. Turriziani, Measuring habituation to stimuli: The italian version of the sensory habituation questionnaire, PloS one 19 (2024) e0309030.

[26] N. E. Adler, T. Boyce, M. A. Chesney, S. Cohen, S. Folkman, R. L. Kahn, S. L. Syme, Socioeconomic status and health: the challenge of the gradient., American psychologist 49 (1994) 15.

[27] A. F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, Communication methods and measures 1 (2007) 77–89.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proc. of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proc. of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[30] G. Geigle, A. Jain, R. Timofte, G. Glavaš, mblip: Efficient bootstrapping of multilingual vision-llms, in: Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR), 2024, pp. 7–25.

[31] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35.

[32] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Grammarly in order to: Paraphrase and reword and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.