The "r" in "woman" stands for rights.
Auditing LLMs in Uncovering Social Dynamics in Implicit Misogyny

Arianna Muti[1], Chris Emmery[2],
Debora Nozza[1], Alberto Barrón-Cedeño[3], Tommaso Caselli[4]
[1]Computing Sciences, Bocconi University; [2]CSAI, Tilburg University;
[3]DIT, University of Bologna; [4]CLCG, University of Groningen
Correspondence: arianna.muti@unibocconi.it

Abstract

Persistent societal biases like misogyny express themselves more often implicitly than through openly hostile language. However, previous misogyny studies have focused primarily on explicit language, overlooking these more subtle forms. We bridge this gap by examining implicit misogynistic expressions in English and Italian. First, we develop a taxonomy of social dynamics, i.e., the underlying communicative intent behind misogynistic statements in social media data. Then, we test the ability of nine LLMs to identify the social dynamics as a multi-label classification and text span selection: first LLMs must choose social dynamics given a prefixed list, then they have to explicitly identify the text spans that triggered their decisions. We also investigate the extent of using different learning settings: zero and few-shot, and prescriptive. Our analysis suggests that LLMs struggle to follow instructions and reason in all settings, mostly relying on semantic associations, recasting claims of emergent abilities.

1 Introduction

In an era dominated by digital communication and social networks, the widespread presence of online misogyny makes online spaces unsafe, perpetuating stereotypes and social injustice. Studies suggest that up to 58% of women have experienced technology-facilitated gender-based violence.[1] Misogyny is not just an individual attitude or prejudice, it is embedded in social structures, cultural norms, institutions and everyday interactions that make up the *social dynamics* between these systems and women. In this context, social dynamics refer to the ongoing patterns of interaction, power exchange, and meaning-making that

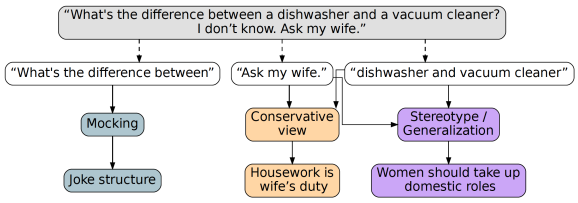

Figure 1: Our framework. Text spans (top) evoke social dynamics (middle), which reflect underlying misogynistic assumptions (bottom).

regulate how women are perceived, treated, and positioned in society. These dynamics often carry implicit communicative content. Examining them can reveal how power relations between genders are maintained, how individuals and groups internalize or reproduce misogynistic ideas, how the gender roles and social expectations shape behaviors toward women, and how norms about masculinity contribute to misogynistic attitudes.

Misogyny and hate speech (HS) datasets more broadly have a relatively low prevalence of *implicitly* hateful content, due to their reliance on explicit keywords during data collection. Ocampo et al. (2023) reports statistics for seven popular HS datasets, revealing that explicit hate occurs, on average, three times more often than implicit hate, while Caselli et al. (2020) show that in their AbuseEval dataset, the percentage of explicit hate is twice that of implicit hate. Specific to misogyny, Muti and Barrón-Cedeño (2022) show that in datasets for English, Italian, and Spanish, the most frequent tokens are swear words, proving their degree of explicitness. However, **focusing primarily on explicit content overlooks the socially pervasive forms of implicit hate, which often evade automated detection and moderation systems yet significantly contribute to the normalization of harmful stereotypes**. Implicit hate operates through euphemism, presupposition, and cultural reference, and its subtlety makes it more likely to

[1] https://www.unwomen.org/en/articles/faqs/digital-abuse-trolling-stalking-and-other-forms-of-technology-facilitated-violence-against-women

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5462–5479
November 4-9, 2025

be perceived as socially acceptable, thereby reinforcing biased social dynamics towards women, under the guise of neutrality.

To effectively capture misogyny, LLMs, like humans, must not only grasp the surface meanings of texts, but also their underlying communicative intention and social knowledge (Choi et al., 2023). While LLMs can often retrieve surface-level associations (first-order meaning), they lack robust metarepresentational awareness needed to infer speaker intent, power dynamics, or social positioning (second-order meaning). The sentence in Fig. 1 illustrates how the joke relies on the implied misogynistic assumption that women are confined to traditional domestic roles.

In this paper, we propose a novel framework informed by feminist and gender studies to annotate implicitly misogynistic social media posts based on the social dynamics they evoke. We then evaluate the capabilities of seven models in English and nine models in Italian in a zero, few-shot, and prescriptive settings to identify social dynamics, along with their corresponding text spans.

We find that LLMs struggle to correctly identify the social dynamics underlying misogynistic statements and the corresponding text span, failing to demonstrate robust reasoning abilities required to effectively interpret these complex social cues without further fine-tuning or knowledge augmentation.

Contributions 1) We create a literature-grounded taxonomy of social dynamics occurring in implicit misogyny (§ 3). 2) We enrich two existing corpora, SBIC (Sap et al., 2020) and ImplicIT-Mis (Muti et al., 2024a) with annotations based on the taxonomy. 3) We audit a total of nine LLMs for their capabilities in identifying the implied social dynamics and their corresponding text span(s). 4) We assess the LLM preference for free-text wrt category selection.

2 Background & Related Work

The meaning and usage of the term **misogyny** have expanded far beyond the original definition; i.e. the hatred of women (Wrisley, 2023). In our work, we intend misogyny as a property of social environments in which women perceived as violating patriarchal norms are "kept down" through hostile or benevolent reactions from men, other women, and social structures (Lopes, 2019; Barreto and Doyle, 2023). As any other forms of hate, misogyny can be expressed explicitly or in a more veiled manner. In the latter case fall instances which are harder to understand for humans —potentially giving rise to disagreements in the annotation phase (Hartvigsen et al., 2022; Yin and Zubiaga, 2022)— and harder to detect for LLMs. Early distinctions on the degree to which hateful content is expressed considered only a binary set (explicit *vs* implicit), where explicitness is defined as unambiguous in its potential of being hateful (Waseem et al., 2017). Ocampo et al. (2023) distinguish between *explicit*, *implicit* and *subtle* HS, grounded in 18 properties of implicitness based on linguistic features, such as sarcasm, figurative language, exaggeration and inferences. While implicit HS goes beyond the literal meaning, in subtle HS there is still a literal meaning, presented in an elusive way. We refer to *implicit misogynous language* when misogyny is expressed through coded or indirect language by means of linguistic devices such as irony, euphemisms, stereotypes, inferences and metaphorical or figurative language (Wiegand et al., 2021).

Datasets designed to address the explicitness degree of hateful messages are rare (Kennedy et al., 2018; Botelho et al., 2021; Caselli et al., 2021; ElSherief et al., 2021). Most current work on **implicit HS** is based either on the re-annotation of existing datasets (Caselli et al., 2020; Wiegand et al., 2021; Ocampo et al., 2023) or on using machine-generated examples (Hartvigsen et al., 2022). Focusing on misogyny, the SOCIAL BIAS INFERENCE CORPUS (SBIC) (Sap et al., 2020), a collection of $10k$ instances of biased statements against minority groups in English, contains the largest group of misogynistic messages: $3.7k$. For Italian, Muti et al. (2024a) developed ImplicIT-Mis, the only dataset with 1,2k implicit misogynous instances. For a complete overview on datasets focused on misogyny see Abercrombie et al. (2023) and Appendix C.

LLMs have introduced a paradigm change in NLP. Their availability has been accompanied by claims about "emerging abilities" (Wei et al., 2022). Recent work has proposed distinguishing the acquisition of competencies in LLMs either as *abilities*; i.e. the capacity to solve a task absent in smaller models as an effect of the size of the models themselves, or *techniques*; i.e. the beneficial effect of different prompting methods that are ineffective in smaller models (Lu et al., 2023). The experiments conducted by Lu et al. (2023) —using zero-shot settings on multiple tasks— show that an ability

Social Dynamics	Description	References	Example
1. Derogatory treatment and belittling of emotions	Demeaning or diminishing women's feelings or experiences, often by belittling their emotions	(Guest et al., 2021)	*she's not depressed she just needs more d*
2. Man-dominated power structure	Situations where men have control or authority over women's decisions, reflecting a power imbalance	(Jane, 2016; Wrisley, 2023)	*women go to the club without their partners to cheat on them*
3. Conservative limitations to women's freedom	Conservative views which limit women's freedom, including criticism for not conforming to traditional roles and promotion of "natural order", abstinence, "pro-life" values	(Siapera, 2019)	*she wastes time on internet instead of being a good wife*
4. Beauty standards expectations	Any expectations on beauty standards, including rejection of self-defined expressions of beauty and appearance	(Amundsen, 2019)	*real beauty is something else, not this fake plastic*
5. Mocking	Any ridiculing or humiliating expression based on jokes, sarcasm, and irony; often appears with offensive terms	(Flick, 2020)	*"ooh, my life is meaningless if I cannot show my tits"*
6. Stereotyping, generalization, prejudices	Oversimplified beliefs about women ignoring individual differences, including stereotypes and unsubstantiated assumptions	(Ging, 2019)	*women are always naked on social media*
7. Whataboutism	A diversion tactic that derails focus from women's issues by raising counterpoints like male victimization or anti-feminist narratives	(Ging, 2019)	*what about violence vs men?*
8. Double standards	Behaviors judged differently based on gender, usually to the detriment of women	(Endendijk et al., 2020)	*it's unattractive when girls act like ghetto*
9. Victim blaming	Especially in sexual assault contexts, when the victim is blamed or held responsible for the attack	(Whatley, 1996)	*she shouldn't have drunk so much*
10. Aggressive and violent attitude	Any threat or hostile behavior posed to women	(Vergès and Thackway, 2022)	*she should be given 2,000 volts*
11. Dismissal of feminism or neosexism	Denying gender inequality or patriarchy, refusing gendered language, or attacking feminists; equating feminism with misandry or extremism	(Siapera, 2019; Muti et al., 2025)	*patriarchy doesn't exist*
12. Sexual objectification	Reducing a person to physical or sexual attributes, ignoring human qualities; opposes the empowerment view of sexuality	(Ging, 2019)	*fresh meat*
13. Centrality of gender distinction	Emphasis on binary gender identity based solely on biological sex; includes heteronormative assumptions	(Fosbraey and Puckey, 2021)	*Born with Dick = Man; Born with Vag = Woman*

Table 1: Overview of the social dynamic categories, their descriptions, corresponding references, and examples taken from our corpora.

such as reasoning is an effect of prompting techniques (e.g., instruction-tuning or in-context learning), rather than an emergent ability. In our paper, we follow a similar experimental setting, where we investigate the behavior of LLMs when it comes to high-level tasks such as classifying and explaining the underlying societal assumptions of implicit misogynistic messages in the form of multi-label classification. The proposed framework serves as a probing method to reveal the social and linguistic knowledge internalized by LLMs during pretraining.

3 Social Dynamics

Understanding and addressing misogyny requires more than identifying overtly harmful language —it demands a deeper analysis of the social structures and assumptions that sustain gender-based inequality. This section introduces a taxonomy of social dynamics developed through the lens of feminist theory and gender studies (Wrisley, 2023; Ramati-Ziber et al., 2019; Srivastava et al., 2017; Lopes, 2019; Kellie et al., 2019; Bergh and Brandt, 2023). Rather than focusing on the linguistic manifestation of misogyny, the taxonomy aims to make

explicit the implicit, unspoken assumptions that underlie misogynistic behavior. This is the main motivation for referring to them as "social dynamics": they describe the manifestation of underlying interactions, attitudes, and behaviors within groups of people (Bannester, 1969). The taxonomy offers a structured framework to classify these underlying dynamics, as summarized in Table 1, which contains the categories, the description taken from literature and an example from our corpora.

We develop this taxonomy through a mixed-method approach, combining bottom-up (inductive) and top-down (deductive) approaches to ensure both empirical grounding and theoretical rigor. In the bottom-up approach, social dynamics emerged from direct observation of data, following grounded theory principles (Glaser and Strauss, 1967) to identify recurring misogynistic behaviors. For the top-down approach, we searched for feminist and gender-based literature that discusses the identified social dynamics to extract underlying explanatory concepts (column Description in Table 1). These concepts were then used to refine the initial categories of our taxonomy, ensuring that the classification was grounded in established theoretical frameworks.

3.1 Data

After defining a structured taxonomy of social dynamics, this study applies it to the annotation of two datasets: the ImplicIT-Mis corpus for Italian (Muti et al., 2024a) and SBIC+ for English (Sap et al., 2020; Muti et al., 2024a). **ImplicIT-Mis** consists of 1,120 Facebook comments that were direct replies to either women-related news articles or to posts on public pages of communities known to tolerate misogyny. An exploration of the top 20 TF-IDF-weighted keywords indicates the lack of any slurs and taboo words, confirming the validity of the corpus for implicit misogyny. The SOCIAL BIAS INFERENCE CORPUS (**SBIC**) (Sap et al., 2020) contains more than $150k$ structured annotations of social media posts to explore the subtle ways in which language can reflect and perpetuate social biases and stereotypes about a thousand demographic groups. To focus exclusively on implicit misogyny, we retained 2.4k messages from the version filtered by Muti et al. (2024a), who selected posts targeting "women" or "feminists" and removed instances with explicit keywords.

3.2 Annotation

Using the defined taxonomy, we enrich 400 instances from the ImpliciIT-Mis and 500 from the SBIC corpora with an annotation layer that targets the social dynamics category (see Table 1), as well as the corresponding text span. The annotation on ImpliciIT-Mis was performed by three Italian linguists, while the SBIC subsection was annotated by one native speaker (from the US) and two Italian linguists, highly proficient in English. The six linguists are experts in gender-based issues. After a training session of one hour, where annotators could ask questions and discuss edge cases, we ran a pilot annotation on 50 instances. The inter-annotator agreement on all social dynamics categories (averaged Cohen's kappa) for English and Italian is 0.460 and 0.339 respectively, showing moderate and fair agreement. This reflects the inherent difficulty of the task, which involves assigning multiple categories and making subjective judgments. Given the high subjectivity of the task and the expertise of the annotators, cases of disagreement have been considered as different perspectives rather than errors, under the lenses of perspectivism (Cabitza et al., 2023) and human label variation (Plank, 2022).

In a second phase, annotators discussed difficult cases, resolving all disagreements collaboratively. In this session, annotators had the opportunity to ask questions to the designer of the Social Dynamic taxonomy[2], enhancing their understanding of the task. In the third phase, annotators proceeded with the annotation of the remaining instances separately. Fig. 2 shows the label distribution for English and Italian. The comparative analysis of implicit misogyny across English and Italian discourse reveals distinct cultural patterns in the expression of implicit misogyny. English data is characterized by a higher prevalence of mocking (approximately 22%) and rejection of feminism (around 12%), suggesting a discourse that frequently employs ridicule and ideological opposition to feminist principles. In contrast, Italian data exhibits greater emphasis on conservative limiting (14%) and beauty standard expectations (12%), indicating a more traditional and appearance-focused manifestation of misogyny. Although both languages show a high prevalence of derogatory and belittling remarks and stereotype generalization, English shows a markedly higher occurrence of

[2]The taxonomy designer is one of the authors.

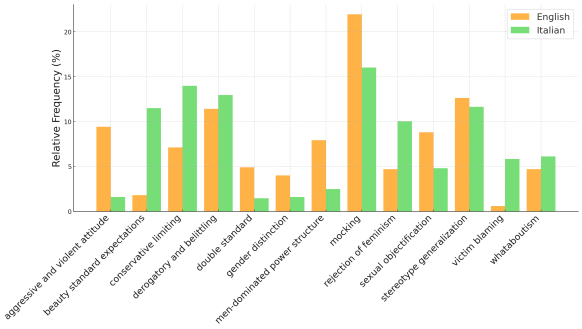

Figure 2: Distribution of social dynamics across English and Italian data.

aggressive and violent attitudes and sexual objectification, whereas Italian discourse relies more heavily on socially ingrained norms and aesthetic expectations. These variations suggest that implicit misogyny in English tends to be more direct and confrontational, whereas in Italian it is more subtly embedded within cultural ideals and gender norms.

For each category labeled by annotators, we also ask to highlight the text span associated to each category. For instance, in the sentence in Fig. 1, *what's the different between* is associated with MOCKING because of the joke structure, while *dishwasher, vacuum cleaner* and *ask my wife* are associated with CONSERVATIVE LIMITING and STEREOTYPE GENERALIZATION. For the inter-annotator agreement of spans we compute the longest common sequence of overlapping characters. The results is 0.634 for English and 0.426 for Italian.

At the end we aggregate all labels identified by annotators to obtain the enriched datasets. In line with the perspectivist paradigm, all labels assigned to each instance are preserved as valid. For the text span, in case of agreement on the category, we retain the longest overlapping span, otherwise each proposed span associated with a category is considered valid.

4 Methodology

In our experiments, we evaluate eight open-source decoder-based LLMs—each with the same number of parameters—alongside GPT-4o-mini on two tasks: (a) selecting relevant social dynamics from a predefined list of categories, and (b) identifying the specific text span that prompted the category selection. We assess model performance under zero-shot, few-shot, and prescriptive prompting conditions. For each model, we provide available details on training data and moderation mechanisms, as these factors significantly influence their behavior and results.

4.1 Models

Due to our available computing infrastructure, and Choi et al. (2023) showing that parameters do not correlate with performance, we end up with models with a 7B-parameter size, except for Llama3 which has 8B. For each model, we select its instruction-tuned version. We follow a zero-shot prompting approach without further fine-tuning the models or providing in-context learning methods for the downstream task (Liu et al., 2023). We present the list of selected models with a description of their main characteristics below. For all models we use the version available on the HuggingFace Model Hub.[3]

Llama2 and Llama3 (`Llama-2-7b-chat-hf` and `Llama-3.1-8B-Instruct`): Both are optimized versions of the original LlaMa model (Touvron et al., 2023). The vast majority of the training materials are in English, although some instances of Italian data are present at training time (0.11%). Furthermore, both models have undergone a phase of safety fine-tuning. For our task, this could trigger instances of over-safety, with the model being unable to follow the instructions and thus failing to provide a valid answer (Röttger et al., 2024). Both have been used for Italian and English.

Mistral-v1 and Mistral-v2 (`Mixtral-8x7B-Instruct-v0.1` and `Mistral-7B-Instruct-v0.2`): This is a family of LLMs based on LlaMa2. Details about the dataset used to generate the models are lacking. In our initial experiments, we have observed that both versions of the models respond to Italian prompts. The instruct-based versions of the models do not present any moderation mechanism. We thus expect this model to avoid over-safety and always provide an answer.

Tower (`TowerInstruct-7B-v0`): This is a multilingual model based on LlaMa2. Multilinguality is achieved by training with a multilingual corpus (20 B tokens over 10 languages, including monolingual and parallel data) and a dedicated dataset for translation-related tasks, including paraphrase generation and named-entity recognition, which might be useful in capturing implicit misogyny (Alves

[3] https://huggingface.co

Task		Models							
		LLama2	**LLama3**	**Mistral-v1**	**Mistral-v2**	**Tower**	**Qwen**	**Minerva**	**GPT-4o-mini**
Social Dynamic	EN	0.00	0.045	0.052	0.067	0.046	0.091	-	**0.134**
	IT	0.00	0.155	0.023	0.053	0.043	0.105	0.197	**0.216**
Text span char. overlap	EN	0.00	0.561	0.469	0.588	0.251	**0.676**	-	0.528
	IT	0.00	0.336	0.369	0.543	0.389	**0.721**	0.661	0.284

Table 2: Macro F1 score in zero-shot setting evaluated on annotated data for English and Italian. For the textual spans, we report the longest common subsequence of character overlap only on matching categories. MFC baselines achieve F1 scores of 0.078 (EN) and 0.063 (IT). Random baselines achieve F1 scores of 0.077 (EN) and 0.072 (IT).

et al., 2024). We have selected this model because it explicitly supports Italian.

Qwen (`Qwen2.5-7B-Instruct`) (Yang et al., 2024). Trained on a diverse dataset of 18 trillion tokens, it features multilingual support for over 29 languages, including Italian.

GPT-4o-mini This is a smaller-scale version of OpenAI's GPT-4o architecture, designed to provide a reduced parameter count and lower memory footprint compared to the full GPT-4o model.

LLaMAntino (`LLaMAntino-2-chat-7b-hf-UltraChat-ITA`): This is a language-specific LLM for Italian (Basile et al., 2023), adapted from LlaMa2. LLamaMantino has been trained on the cleaned Italian split of the multilingual Common Crawl's web crawl corpus (Sarti and Nissim, 2022).

Minerva (`Minerva-7B-instruct-v1.0`) (Orlando et al., 2024). It is the first LLM trained from scratch on native Italian texts. It is trained on post-processed web data and curated data, including Wikipedia, EurLex and Gazzetta Ufficiale (law, economics, and politics), and the Gutenberg Project (novels, poetry). The architecture is based on Mistral.

4.2 Prompt Definition

It is a known phenomenon that the specific format of a prompt may result in very different outcomes when applied to LLMs, also in hate speech contexts (Plaza-del arco et al., 2023). To control for this, we investigate the behaviors of the models to follow prompt instructions. We first tested approximately 100 prompts across both languages in zero-shot settings. Our changes mainly involved the use of synonyms and descriptions of how the final output should be structured. In some cases, we opted to use letter-based labels instead of verbal category names. For the Italian models, we have translated the instructions into Italian. We have used an initial set of 50 sentences in English and Italian to analyze the output and decide on the final prompt format. After having identified the *zero-shot* prompt that yielded the most coherent and consistent results, we implemented a *few-shot* setting by providing 13 illustrative examples, one for each social dynamic category. We also experimented with the *prescriptive* setting (Rottger et al., 2022), in which we incorporated the full set of annotation guidelines, including examples, directly into the prompt. We prompt LLaMa2, LlaMa3, Mistral-v1, Mistral-v2, Qwen, Tower and GPT-4o-mini on both English and Italian data. The Italian LLMs, LLaMAntino and Minerva, were prompted only on the Italian data. When running LLaMa, Mistral, and Tower on the Italian data, we use the English prompt. For Qwen, Minerva, and Llamantino we use the Italian prompt. For all models, we set the temperature to 0. When running LLaMa, Tower and Mistral on the Italian messages, we observed a tendency of these models to translate the Italian input or text spans into English. To limit this, we explicitly asked such models not to translate the message in the prompt. Our final prompt instructs the models to select one or more of the 13 social dynamics discussed in § 3 as well as the corresponding text span. The English prompts and the Italian translation are reported in Appendix A. For the prompt design, we took inspiration from Hromei et al. (2023); Lu et al. (2023).

5 Experiments and Results

We perform a quantitative and qualitative analysis of the LLMs outputs in order to analyze *i*) whether the models are able to complete the task properly with respect to the instructions and structure of the output and *ii*) to what extent the models can predict the social dynamic category(ies) and the corresponding text span(s). We compare our results against two baselines: Random and Most Frequent

Model	Few-shot		Prescriptive	
	EN	IT	EN	IT
GPT-4o-mini	0.110	**0.138**	0.115	0.116
Qwen	0.116	0.139	**0.164**	0.176

Table 3: Macro F1 score in few-shot and prescriptive settings.

Class (MFC), which always predict the two most frequent categories, based on the observed average of two annotated categories per instance in both English and Italian.

LLMs output validity Although we explicitly prompt the models to produce output in a specific format, many often fail to do so. Consequently, we perform ad-hoc automatic post-processing to extract the relevant categories and text spans. Nonetheless, some portions of the responses cannot be reliably extracted. Additionally, some models occasionally refuse to respond. We also observe systematic errors in the handling of social dynamics categories; models may alter category names or introduce new ones. To mitigate this, we apply post-processing to map incorrect or unknown categories to the closest valid alternatives whenever possible (see Appendix B for more details).

The percentages of outputs for which we were unable to extract categories and spans (English / Italian) are as follows: LLama2 (59% / 12%), LLama3 (13% / 9%), Mistral-v1 (2% / 5%), Mistral-v2 (17% / 28%), Tower (18% / 14%), Qwen (0% / 3%), Qwen few-shot (0% / 6%), Qwen perspective (24% / 9%), Minerva (— / 0%), GPT-4o-mini (0% / 0%), GPT few-shot (0% / 0%), GPT perspective (0% / 0%), Llamantino (— / 82%). Llamantino was excluded from the evaluation due to a high proportion of missing output categories, largely caused by the model's tendency to respond in a dialogue-oriented manner rather than following the instructed format. The high percentages of missing outputs for LLama2 are due to the model's frequent refusals to answer, a result of its overly cautious safety mechanisms, as previously observed in Röttger et al. (2024).

Human Evaluation Table 2 reports the Macro F1 score against the human annotations for the social dynamics categories and the text spans. Appendix D shows per-category performance. Overall performance is generally low, reflecting the inherent difficulty of the task. The large number of categories and the subjective nature of the classification contribute to this complexity, mirroring the challenges observed during human annotation. This is further evidenced by the performance of the Random and MFC baseline models, which produce similar results (see Table caption).

Models achieve higher performance with Italian data. GPT-4o-mini achieves the best score in English, although as low as 0.115, and in Italian with 0.216. In English, models correctly predict only the very few explicit cases of SEXUAL OBJECTIFICATION and AGGRESSIVE AND VIOLENT ATTITUDE that bypassed the filtering process, and MOCKING when it has a clear joke structure: *what's the difference between X and a woman?*. In Italian, all categories are correctly predicted at least once, with SEXUAL OBJECTIFICATION being the most frequent across all models, followed by VICTIM BLAMING and MOCKING. Concerning the text spans, Qwen is the best at capturing the longest common sequences in both languages. Since our zero-shot approach yielded low results, we explore few-shot and prescriptive learning setting with the best models, namely Qwen and GPT-4o-mini (Table 3). Although these new learning paradigms allow us to leverage contextual learning, performance decreases with GPT-4o-mini (by 17.9% in English and 36.1% in Italian), while it increases with Qwen (by 80.2% in English and 31.4% in Italian), achieving the highest performance in the prescriptive setting in English.

In all settings, we observe the **models' tendency to select the social dymanic category based on semantic associations instead of the intended meaning of the message**. For instance, the sentence *non è colpa loro se sono stupide - You can't blame them for being stupid*, gets labeled as victim blaming, in Italian "colpevolizzazione della vittima", most likely for the semantic association between *colpa (blame)* and *colpevolizzare (blaming)*. The same happens in English, in which *the next girl to reject me* gets associated with REJECTION OF FEMINISM. We compute the overlapping tokens with and without stemming, after having excluded stopwords, in English and Italian (see Table 4).

Free-text vs. multiple choice: what do LLMs prefer? We select the best-performing model overall, GPT-4o-mini, and prompt it to elicit the implying communicative intent before categorizing the social dynamics. We manually observe 100 random instances from both the English and Ital-

Language	Model	Setting	Overlap_tokens	Overlap_stem	Words (frequency)
EN	GPT-4o-mini	zero-shot	45	54	male (15), women (10), feminism (6), sexual (3), reject (3), sexism (3), violent (2), gender (2), beauty (2), generalization (2), freedom (1), emotion (1), power (1), structure (1), objectification (1), standard (1)
EN	GPT-4o-mini	few-shot	40	46	women (12), male (10), feminism (6), gender (5), sexism (3), sexual (3), beauty (2), violent (2), blame (2), power (1), emotion (1), reject (1), respect (1), standard (1)
EN	GPT-4o-mini	guidelines	32	50	women (14), feminism (7), male (7), gender (4), sex (4), beauty (4), violence (2), power (2), freedom (2), victim (1), blame (1), body (1), objectification (1)
IT	GPT-4o-mini	zero-shot	10	23	femmin* (9), uomini (3), bellezza (3), donne (2), misura (1), aspetto (1), generalizzazione (1), rispetto (1), stereotipi (1), vittima (1)
IT	GPT-4o-mini	few-shot	22	29	donne (12), bellezza (5), generalizzazione (3), rispetto (2), violento (1), libertà (1), aspetto (1), trattamento (1), uomini (1), vittima (1), sessuale (1)
IT	GPT-4o-mini	guidelines	19	26	donne (13), bellezza (5), aspetto (2), rispetto (2), vittima (1), limite (1), trattamento (1)

Table 4: Overlap between model-predicted category labels and lexically associated words by language, model, and setting.

Multiple choice	✓	✓	✗	✗
Free-text	✓	✗	✓	✗
EN	57	7	19	17
IT	46	15	16	23

Table 5: Number of instances (out of 100) across the four evaluation conditions.

ian datasets to evaluate whether LLMs are more effective at representing social knowledge through structured social dynamics categories or free-text implied assumptions and if there is a correspondence between the explanation and the category. We conduct a binary evaluation based on four conditions: (a) at least one social dynamic and the explanation are correct; (b) at least one social dynamic is correct while the explanation is incorrect; (c) the social dynamic(s) is/are incorrect while the explanation is correct; and (d) both are incorrect. Table 5 shows the results. This comparison not only informs us about model evaluation, but it also sheds light on how LLMs operationalize complex, context-dependent social meaning in practice.

In English, in setting (a), we found instances containing mostly explicit misogyny in the form of traditional gender roles, violence, and sexual objectification. These cases were unexpected, as the dataset was labeled as containing implicit content. It is therefore reasonable to assume that these more explicit instances are easier for the model to detect. An exception of an implicit text falling in setting (a) is *what does the e in woman stand for? equal*, where GPT-4o-mini responds that *the concept of equality is not genuinely represented in the term "woman"*, correctly mapping it to REJECTION OF FEMINISM. However, most of subtle instances fall in (b), where the category is correctly identified because of words associations, but the underlying meaning is not correct. E.g.: *i said hi to a feminist today the court date is set for november 8th* implies that feminists are prone to falsely accusing men of sexual harassment or overreacting to benign interactions. However, the model lacks adequate social knowledge. In setting (c), seven instances were judged incorrect but could reasonably be considered valid under alternative interpretations. For instance, *i just don't find women funny* has been labeled as MEN-DOMINATED POWER STRUCTURE, GENDER DISTINCTION, but STEREOTYPE GENERALIZATION predicted apply as well. Moreover, in three cases, the predicted category is wrong, but the implied assumption contains the term of the correct category. For instance, the category predicted is DEROGATORY AND BELITTLING, but the implied assumption *The text diminishes the mother's feelings of grief and panic by presenting a violent and mocking scenario regarding the loss of her child* implicitly contain the two correct categories: MOCKING, VIOLENT ATTITUDE. Setting (d), the one where both the multiple choice and free-text approaches fail, mainly contains puns, sarcasm and complex reasoning, as in *a woman saw me at a condom machine the other day... she said "you're optimistic" then i said "no, i'm just stronger than you."*, where the punchline is a reference to using physical force to obtain sex. In Italian, in setting (a) most instances are clear examples of VICTIM BLAMING. Overall, in settings (b) and (c) explanations are very general, making it unclear whether the model has the required social knowledge to identify the problematic passage. For instance, in the sentence *Like Elodie she stood on the freeway*, GPT-4o-mini guesses the category DEROGATORY AND

BELITTLING, but the explanation lacks the motivation. The sentence implicitly links the unnamed female and Elodie, an Italian singer, to prostitution through the reference to standing on the freeway, an euphemism for street prostitution. However, the explanation does not mention it: *The text implies a judgment or criticism of Elodie's behavior or situation, possibly suggesting that it is inappropriate or undesirable*. In setting (d), sarcasm is common, often implying the opposite of what is said, and includes Italian slang (e.g., *menne* for breast), which highlights the limitations of GPT-4o-mini in understanding regional or informal variations of Italian.

Although there is a strong association between correct categories and free-text explanations, the latter are more accurate than the predicted categories, especially in English, suggesting a model preference for capturing underlying reasoning over adhering strictly to the classification schema. However, this does not always guarantee precision: free-texts tend to be very general and high-level forms of misogyny do not emerge clearly in free-text, reinforcing the complementary role of structured classification. For instance, comments reflecting men-dominated power structure are described in free-text as "dependent on men", "power imbalance where men have authority over women's identities" and other ways that lack the specificity of the structured label and make the evaluation harder. For these reasons, the ideal scenario is to combine both forms—free-text explanations and categorical labels—as they offer complementary strengths in interpretability and accuracy.

6 Conclusion

We audit nine LLMs to evaluate their ability to identify the social dynamics encoded in implicit misogynistic messages in Italian and English, and the corresponding text span. To this end, we propose a literature-grounded taxonomy of social dynamics that occur in implicit misogynous statements.

Overall, models struggle to complete such a task which requires complex reasoning about meaning, and rely more on surface-level semantic associations. We also compared category selection vs free-text explanations, revealing that while explanations more often capture the underlying reasoning, they lack the precision of structured labels, highlighting the importance of combining both approaches for more accurate and interpretable model evaluation. In either case, **the more social knowledge is required to understand the underlying meaning of the message, the likelier models fail.** While this supports previous findings on recasting claims on emerging abilities of LLMs (Lu et al., 2023), it also indicates that LLMs have limited understanding of implied societal assumptions encoded in messages. Hence, additional training/tuning or knowledge-augmentation approaches are needed. We leave the use of external knowledge for future work.

Limitations

A key limitation of this work is our exclusive use of smaller language models, specifically those in the 7 billion parameter range and the cost-efficient GPT-4o-mini. While larger models often offer superior performance on a wide array of tasks, we intentionally constrained our experimentation to lightweight alternatives. This choice was driven by two core principles: fostering inclusive research practices and minimizing computational overhead. First, limiting our model choices supports broader accessibility and reproducibility within the research community. High-performance computing resources and the financial means to access large-scale proprietary models remain concentrated among well-funded institutions and corporations. By focusing on smaller, more accessible models, we aim to lower the barrier to entry for independent researchers, educators, and groups in resource-constrained settings. Second, the emphasis on reducing computational overhead reflects a commitment to environmentally and economically sustainable research. Nonetheless, this design choice inevitably limits the generalizability of our findings to more capable, large-scale models.

Ethical Considerations

Measures were taken to alleviate and monitor the mental health of the annotators. The annotators were in constant communication with one of the authors and they had a session in which they could express their concerns, however, none were raised. Moreover, they were compensated beyond the average national wage (10 EUR per hour).

Acknowledgements

Arianna Muti's and Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Arianna Muti and

Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Rikke Amundsen. 2019. *Cruel Intentions and Social Conventions: Locating the Shame in Revenge Porn*, pages 131–148. Springer International Publishing, Cham.

E. Michael Bannester. 1969. Sociodynamics: An integrative theorem of power, authority, interfluence and love. *American Sociological Review*, 34(3):374–393.

Manuela Barreto and David Matthew Doyle. 2023. Benevolent and hostile sexism in a shifting global context. *Nature reviews psychology*, 2(2):98–111.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Robin Bergh and Mark J. Brandt. 2023. Generalized prejudice: Lessons about social power, ideological conflict, and levels of abstraction. *European Review of Social Psychology*, 34(1):92–126.

Austin Botelho, Scott Hale, and Bertie Vidgen. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1896–1907, Online. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joyce J. Endendijk, Anneloes L. van Baar, and Maja Deković. 2020. He is a stud, she is a slut! a meta-analysis on the continued existence of sexual double standards. *Personality and Social Psychology Review*, 24(2):163–190. PMID: 31880971.

Caterina Flick. 2020. *The Legal Framework on Hate Speech and the Internet Good Practices to Prevent and Counter the Spread of Illegal Hate Speech Online: Good Practices to Prevent and Counter the Spread of Illegal Hate Speech Online*.

Glenn Fosbraey and Nicola Puckey. 2021. *Misogyny, Toxic Masculinity, and Heteronormativity in Post-2000 Popular Music*.

Debbie Ging. 2019. *Bros v. Hos: Postfeminism, Antifeminism and the Toxic Turn in Digital Gender Politics*, pages 45–67. Springer International Publishing, Cham.

Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, Chicago.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Claudiu Daniel Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. 2023. Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy. CEUR.org.

Emma Jane. 2016. *Misogyny Online: A Short (and Brutish) History*.

Dax J. Kellie, Khandis R. Blake, and Robert C. Brooks. 2019. What drives female objectification? an investigation of appearance-based interpersonal perceptions and the objectification of women. *PLoS One*, 14(8):e0221388.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and et al. 2018. Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale.

Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.

Kätriin Kukk, Danila Petrelli, Judit Casademont, Eric J. W. Orlowski, Michal Dzielinski, and Maria Jacobson. 2025. BiaSWE: An expert annotated dataset for misogyny detection in Swedish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 307–312, Tallinn, Estonia. University of Tartu Library.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Filipa Melo Lopes. 2019. Perpetuating the patriarchy: Misogyny and (post-)feminist backlash. *Philosophical Studies*, 176(9):2517–2538.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

Arianna Muti and Alberto Barrón-Cedeño. 2022. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, Dublin, Ireland. Association for Computational Linguistics.

Arianna Muti, Sara Gemelli, Emanuele Moscato, Emilie Francis, Amanda Cercas Curry, Flor Miriam Plaza-del Arco, and Debora Nozza. 2025. Blue-haired, misandriche, rabiata: Tracing the connotation of 'feminist(s)' across time, languages and domains. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 299–311, Vienna, Austria. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024a. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. 2024b. PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia. ELRA and ICCL.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Leeat Ramati-Ziber, Nurit Shnabel, and Peter Glick. 2019. The beauty myth: Prescriptive beauty norms for women reflect hierarchy-enhancing motivations leading to discriminatory employment practices.

Journal of Personality and Social Psychology, pages 1–27.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *ArXiv preprint 2203.03759*.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2024. Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 427–452, Bangkok, Thailand. Association for Computational Linguistics.

Eugenia Siapera. 2019. *Online Misogyny as Witch Hunt: Primitive Accumulation in the Age of Technocapitalism*, pages 21–43.

Kalpana Srivastava, Suprakash Chaudhury, Pookala Bhat, and Samiksha Sahu. 2017. Misogyny, feminism, and sexual harassment. *Industrial Psychiatry Journal*, 26:111.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Françoise Vergès and Melissa Thackway. 2022. *A Feminist Theory of Violence: A Decolonial Perspective*. Pluto Press.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Mark A. Whatley. 1996. Victim characteristics influencing attributions of responsibility to rape victims: A meta-analysis. *Aggression and Violent Behavior*, 1(2):81–95.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Samantha Pinson Wrisley. 2023. Feminist theory and the problem of misogyny. *Feminist Theory*, 24(2):188–207.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

A Prompts

A.1 Zero-shot Settings

Setting	Prompt
Zero-Shot EN	Choose the social dynamics implied in the text provided between triple quotes. Report also the part of the text that triggered your choice. Do not provide further explanation. Choose the social dynamics from the options: "victim blaming", "derogatory treatment or belittling of emotions", "male-dominated power structure", "expectations with respect to beauty standards", "conservative view that limits women's freedom", "mockery", "stereotyping, generalization, unfounded assumptions, prejudice", "whataboutism", "double standards", "aggressive and violent attitude", "dismissal of feminism or neo-sexism", "sexual objectification", "centrality of gender distinction". Answer in the format [social dynamic: "text span"] "'text"'
Zero-Shot IT	Scegli le dinamiche sociali implicite nel testo tra virgolette triple. Riporta la parte di testo responsabile della tua scelta. Scegli una o più tra le seguenti opzioni: "colpevolizzazione della vittima", "trattamento dispregiativo o sminuente delle emozioni", "struttura di potere dominata dagli uomini", "aspettative rispetto agli standard di bellezza", "visione conservatrice che limita la libertà delle donne", "derisione", "stereotipi, generalizzazioni, ipotesi infondate, pregiudizi", "benaltrismo", "due pesi due misure", "atteggiamento aggressivo e violento", "rifiuto del femminismo o del neosessismo", "oggettificazione sessuale", "centralità della distinzione di genere". Non fornire ulteriori spiegazioni. Rispondi nel formato [dinamica sociale:"parte di testo"] "'testo"'

Table 6: Text used to prompt the model in English and Italian in zero-shot settings.

A.2 Few-shot Settings

Prompt: Read the text between triple quotation marks.
1. Think about the implicit assumption of the text, then choose one or more social dynamics that the implicit assumption evokes. "derogatory or belittling treatment of emotions", "male-dominated power structure", "expectations with respect to beauty standards", "conservative view that limits women's freedom", "mocking", "stereotypes, generalizations, unfounded assumptions, prejudices", "whataboutism", "double standards", "aggressive and violent attitude", "rejection of feminism or neo-sexism", "sexual objectification", "victim blaming", "centrality of gender distinction".
2. Select the part of the text associated with each chosen category.
3. Respond in the format [category: "text span"].

Examples:

Text: only lesbians get their nips pierced ew [Derogatory treatment and belittling of emotions: "only lesbians get their nips pierced ew"], [Expectations with respect to beauty standards:"nips pierced ew"]

Text: I've seen some very beautiful white women and nothing sickens me more than to see them pull up at the mall, grocery store, or gas station and think damn she's beautiful then to see her open the back door and walk in with an oreo baby kinda makes you wish it was still 1940 here in alabama [Man-dominated power structure: "she's beautiful then to see her open the back door and walk in with an oreo baby kinda makes you wish it was still 1940 here in alabama"]

Text: No girl should be allowed to go partying without their men [Conservative view that limits women's freedom: "No girl should be allowed to go partying without their men"]

Text: I like my wife as I like my whisky: 30 years old and locked in the cellar [Mocking: "I like my wife as I like my whisky: 30 years old and locked in the cellar"]

Text: All I know is that every girl my age has had at least 5 sexual partners and I'm still a virgin. Unless of course it's the same guys that fuck all girls which is a bit comforting otherwise just kill me. [Stereotyping, generalization and unfounded prejudices: "every girl my age has had at least 5 sexual partners"]

Text: what about violence against men? [Whataboutism: "what about violence against men?"]

Text: it's unattractive when girls act like ghetto [Double standards: "it's unattractive when girls act like ghetto"]

Text: they blame it on the alcohol but claim the truth come out when they drunk.. [Victim blaming:"they blame it on the alcohol"]

Text: she should be given 2,000 volts [Aggressive and violent attitude:"she should be given 2,000 volts"]

Text: patriarchy doesn't exist [Rejection of feminism or neo-sexism: "patriarchy doesn't exist"]

Text: breh it's a polygamist town in arizona if u wanna have bitches on deck move there [Sexual objectification: "it's a polygamist town in arizona if u wanna have bitches on deck"]

Text: Born with Dick = Man; Born with Vag = Woman [Centrality of gender distinction: "Born with Dick = Man; Born with Vag = Woman"]

For Italian, we took the Italian examples from the prescriptive prompt.

A.3 Prescriptive Settings

Follow the instructions for the annotations.

Read carefully the list of the social dynamics along with the examples.
Select any social dynamic category that applies to each text instance.
Select the span of the text that activates the social dynamic(s) you have selected.
If the social dynamic is expressed by the entire text, select the whole text.

Social Dynamics in Misogynistic Content
The proposed taxonomy of social dynamics is grounded in feminist literature and gender studies and focuses on making explicit the underlying assumption(s) that trigger a misogynous interpretation rather than the linguistic manifestation(s) of misogyny.
For instance, the misogynistic category of "body shaming" can be seen as a manifestation of a social dynamic for which there are expectations rooted in a society with respect to beauty standards.
This is the main motivation for referring to these categories as "social dynamics":
a social dynamic describes the manifestation of underlying interactions, attitudes, and behaviors towards women as they occur in a societal context.
Our social dynamics taxonomy can be seen as categorizations of misogynistic implied assumptions.

Consider the following example sentence:

"True beauty is something else, not this trashy plastic"

This text implies that women who undergo plastic surgery are not beautiful. This implication can be traced back to the social dynamics of "conservative view that limits women's freedom" and "expectations wrt beauty standards", because:

conservative view that limits women's freedom: women should not undergo plastic surgery
expectations wrt beauty standards: natural beauty is the acceptable standard for women

The annotation is a three step process:
Identify the underlying communicative content (implication) of the text
Assign any social dynamic that applies from the list
Associate each social dynamic with the corresponding text span or the entire message.

Below you will find the list of social dynamics, along with definitions and examples for each category in IT and EN. Consider only the English (or Italian) examples for this round of annotations.
Spans are highlighted. Please keep in mind that these sentences contain more than one social dynamic, but I have selected only one instance for clarity's reasons.

Social Dynamics

1. Derogatory treatment and belittling of emotions: Demeaning or diminishing women's feelings or experiences, often by belittling their emotions or showing disgust for their actions.

IT: non è depressa deve solo prenderne di più
EN: only lesbians get their nips pierced ew

2. Man-dominated power structure: Situations where men have control or authority over women's decisions, reflecting a power imbalance.

IT: Corna assicurate la ragazza a ballare da sola non va, non bisogna dargli troppa libertà sono corna assicurate...
EN: ive seen some very beautiful white women and nothing sickens me more than to see them pull up at the mall , grocery store , or gas station and think damn shes beautiful then to see her open the back door and walk in with an oreo baby kinda makes you wish it was still 1940 here in alabama

3. Conservative view that limits women's freedom: this includes criticism for not conforming to traditional roles or expectations, references to a "natural order", sexual abstinence and "pro-life" values.

IT: guarda questa invece di fare la brava mogliettina cosa fa
EN: No girl should be allowed to go partying without their men

4. Expectations with respect to beauty standards: Any expectations on beauty standards, including the rejection of self-defined expressions of beauty and appearance.

IT: la vera bellezza è un'altra, non questa plasticaccia EN: only lesbians get their nips pierced ew

5. Mocking: Any ridiculing or humiliating expression based on jokes, sarcasm and irony. This category is seen alongside offensive terms used for other categories.

IT: la mia vita non ha senzo se non posso uscirlee
EN: I like my wife as I like my whisky: 30 years old and locked in the cellar

6. Stereotyping, generalization and unfounded prejudices: Oversimplified beliefs about women that do not consider individual differences; it can be stereotypical with respect to the traditional gender role of women or made-up generalizations; it can also be prejudice without evidence.

IT: E SEMPRE ESISTITO CHE ALLE DONNE PIACCIONI I TIPI "STRANI", "CATTIVI" "POSSESSIVI".....I BRAVI RAGAZZI NON SE LE FILA NESSUNA. Poi crescono e passano ad osservare il conto in banca.
EN: All I know is that every girl my age has had at least 5 sexual partners and I'm still a virgin. Unless of course it's the same guys that fuck all girls which is a bit comforting otherwise just kill me.

7. Whataboutism: A diversion tactic that shifts focus and derail conversations from issues affecting women by raising counterpoints about other issues, such as domestic violence against men, child custody, divorce and the feminization of education.

IT: Anche le donne sono violente ! Noi uomini dovremmo farci un esame di coscienza? Col cazzo!!
EN: what about violence against men?

8. Double standards: Behaviors are judged differently based on gender, often detrimental to women. The sentence could be completed by 'men instead. . . '.

IT: Le donne non dovrebbero andare in giro da sole
EN: it's unattractive when girls act like ghetto

9. Victim blaming: Usually in the context of sexual assault or femicide, when the victim is held responsible.

IT: Vittime e carnefici si attraggono. L'educazione del maschio all'affettività va bene, anche se non nelle modalità auspicate. Ma anche la donna ha un lungo percorso da fare per ritrovare se stessa. EN: they blame it on the alchohol but claim the truth come out when they drunk..

10. Aggressive and violent attitude: Any threat or hostile behavior that is posed to women.

IT: TSO
EN: she should be given 2,000 volts

11. Rejection of feminism or neo-sexism: Denying the existence of gender inequality, patriarchy. It includes the refusal of using gendered language (including feminine job titles) and personal attacks against feminists. It also includes positions explicitly against or mocking/denying gender equality or gender issues; statements that feminism is misandry or is compared to Nazism, and that men have to fight against their diminished power in society.

IT: si dice CARABINIERE non CARABINIERA; io ho sempre detto che le donne vogliono l" ugualianza quando e vantaggiosa , e vogliono essere donne quando e conveniente
EN: patriarchy doesn't exist

12. Sexual objectification: The reduction of a person to their physical attributes or sexual appeal, including men and women who perceive sexualized women as lacking certain human qualities such as mental capacity and moral status. This opposes to the view of self-enhancement and sexuality as empowering.

IT: perché vuoi lavorare così figa?
EN: breh it's a polygamist town in arizona if u wanna have bitches on deck move there

13. Centrality of gender distinction: Emphasizes binary views of gender identity based solely on biological sex, including what women and men are supposed to do and how should behave based on their biological sex. Disparaging heteronormativity can also fit in this category.

IT: Più penoso dell'uomo comunista c'è la donna comunista.
EN: Born with Dick = Man; Born with Vag = Woman.

Free-text Read the text between triple quotation marks. 1. Think about the implicit assumption of the text, then choose one or more social dynamics that the implicit assumption evokes. "derogatory or belittling treatment of emotions", "male-dominated power structure", "expectations with respect to beauty standards", "conservative view that limits women's freedom", "mocking", "stereotypes, generalizations, unfounded assumptions, prejudices", "whataboutism", "double standards", "aggressive and violent attitude", "rejection of feminism or neo-sexism", "sexual objectification", "victim blaming", "centrality of gender distinction". 2. Select the part of the text associated with each chosen category. 3. Respond in the format [category: "text span"].

Examples:

Text: only lesbians get their nips pierced ew [Derogatory treatment and belittling of emotions: "only lesbians get their nips pierced ew"], [Expectations with respect to beauty standards:"nips pierced ew"]

Text: I've seen some very beautiful white women and nothing sickens me more than to see them pull up at the mall, grocery store, or gas station and think damn she's beautiful then to see her open the back door and walk in with an oreo baby kinda makes

you wish it was still 1940 here in alabama [Man-dominated power structure: "she's beautiful then to see her open the back door and walk in with an oreo baby kinda makes you wish it was still 1940 here in alabama"]

Text: No girl should be allowed to go partying without their men [Conservative view that limits women's freedom: "No girl should be allowed to go partying without their men"]

Text: I like my wife as I like my whisky: 30 years old and locked in the cellar [Mocking: "I like my wife as I like my whisky: 30 years old and locked in the cellar"]

Text: All I know is that every girl my age has had at least 5 sexual partners and I'm still a virgin. Unless of course it's the same guys that fuck all girls which is a bit comforting otherwise just kill me. [Stereotyping, generalization and unfounded prejudices: "every girl my age has had at least 5 sexual partners"]

Text: what about violence against men? [Whataboutism: "what about violence against men?"]

Text: it's unattractive when girls act like ghetto [Double standards: "it's unattractive when girls act like ghetto"]

Text: they blame it on the alcohol but claim the truth come out when they drunk.. [Victim blaming:"they blame it on the alcohol"]

Text: she should be given 2,000 volts [Aggressive and violent attitude:"she should be given 2,000 volts"]

Text: patriarchy doesn't exist [Rejection of feminism or neo-sexism: "patriarchy doesn't exist"]

Text: breh it's a polygamist town in arizona if u wanna have bitches on deck move there [Sexual objectification: "it's a polygamist town in arizona if u wanna have bitches on deck"]

Text: Born with Dick = Man; Born with Vag = Woman [Centrality of gender distinction: "Born with Dick = Man; Born with Vag = Woman"]

Text: "'text"'

B Models' errors.

In general, we see errors for both social dynamics and text spans predictions. For the social dynamics categories, we observe that models tend to either distort the name of the categories (e.g., "whatsaboutism" instead of "whataboutism") or truncate the name of categories, up to inventing new categories. We handled such cases with post-processing, mapping wrong categories to the closest correct ones when possible. We first used ChatGPT for the mapping and then manually checked. For instance, if the model outputs "stereotype on beauty standards" we map it to "beauty standard expectations". For English, we found a total of 91 made-up categories that could not be mapped, among which "homophobia", "societal critique", and "sexual assault". The model that mostly generates new social dynamics is Mistral-v1 (with 88 made-up categories). In Italian, the unmapped categories are significantly lower, only 10 in total. Regarding the text spans, we observe that in both languages models produce verbose answers by providing an explanation of why the sentence is misogynous. Tower tends to generate new texts more often, rather than extracting the text spans from the message. Italian LLMs tend to translate the original text.

C Complimentary Related Work

Dataset	Scale & Source	Annotation Schema	Annotation Methodology	Language	Reference
Biasly	Movie dialogue subtitles; 10K texts	Binary misogyny flag; continuous severity score; suggested rewrites per instance	Multi-task annotation by domain experts and trained annotators	English	Sheppard et al. (2024)
PejorativITy	1,200 Italian tweets	Word-level pejorativity annotation; sentence-level binary misogyny	Six trained annotators for the pilot; discussion panels	Italian	Muti et al. (2024b)
BiaSWE	450 posts from Swedish forum Flashback	Binary misogyny; misogyny types; severity levels	Expert annotation	Swedish	Kukk et al. (2025)
GerMS-AT	8K forum comments from an Austrian online newspaper	Five-level sexism/ misogyny severity	Annotated by professional forum moderators, with expert-defined guidelines	German (Austrian-dialect)	Krenn et al. (2024)

Table 7: Overview of datasets with misogyny/sexism annotations not covered by Abercrombie et al. (2023).

D Detailed Results on Social Dynamics Categories

Table 8 shows classification performance across categories in zero-shot settings. For EN, the highest F1-scores are achieved in mocking (0.60, GPT), sexual objectification (0.50, GPT), aggressive and violent attitude (0.49, GPT), and beauty standard expectations (0.27, Qwen), indicating that models capture overtly hostile or objectifying language with relatively greater reliability. In contrast, IT shows its strongest performance in beauty standard expectations (0.50, GPT), rejection of feminism (0.47, GPT), derogatory and belittling (0.34, LLaMA3), and stereotype generalization (0.34, GPT). This suggests that while EN classifiers perform better at detecting mocking and objectification, IT classifiers are comparatively stronger in identifying normative or ideological biases.

	EN			IT		
Category	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Tower						
aggressive and violent attitude	0.22	0.14	0.17	0.06	0.30	0.09
beauty standard expectations	0.00	0.00	0.00	0.00	0.00	0.00
conservative limiting	0.00	0.00	0.00	0.00	0.00	0.00
derogatory and belittling	0.00	0.00	0.00	0.00	0.00	0.00
double standard	0.00	0.00	0.00	0.00	0.00	0.00
gender distinction	0.00	0.00	0.00	0.00	0.00	0.00
men-dominated power structure	0.00	0.00	0.00	0.00	0.00	0.00
mocking	0.00	0.00	0.00	0.00	0.00	0.00
rejection of feminism	0.00	0.00	0.00	0.00	0.00	0.00
sexual objectification	0.23	0.43	0.30	0.13	0.50	0.21
stereotype generalization	0.00	0.00	0.00	0.00	0.00	0.00
victim blaming	0.01	0.33	0.02	0.10	0.59	0.16
whataboutism	0.08	0.15	0.10	0.07	0.14	0.10
Qwen						
aggressive and violent attitude	0.38	0.03	0.06	0.25	0.10	0.14
beauty standard expectations	0.26	0.28	0.27	0.31	0.07	0.11
conservative limiting	0.32	0.15	0.21	0.35	0.13	0.19
derogatory and belittling	0.00	0.00	0.00	0.17	0.12	0.14
double standard	0.13	0.10	0.11	0.17	0.11	0.13
gender distinction	0.10	0.07	0.09	0.03	0.44	0.06
men-dominated power structure	0.31	0.35	0.33	0.11	0.33	0.16
mocking	0.57	0.16	0.26	0.34	0.29	0.31
rejection of feminism	0.30	0.30	0.30	0.54	0.26	0.35
sexual objectification	0.59	0.19	0.29	0.17	0.28	0.21
stereotype generalization	0.00	0.00	0.00	0.17	0.16	0.16
victim blaming	0.14	0.17	0.15	0.20	0.12	0.15
whataboutism	1.00	0.09	0.16	0.17	0.03	0.05
Mistral						
aggressive and violent attitude	0.26	0.29	0.27	0.00	0.00	0.00
beauty standard expectations	0.00	0.00	0.00	0.00	0.00	0.00
conservative limiting	0.00	0.00	0.00	0.00	0.00	0.00
derogatory and belittling	0.00	0.00	0.00	0.00	0.00	0.00
double standard	0.00	0.00	0.00	0.00	0.00	0.00
gender distinction	0.00	0.00	0.00	0.00	0.00	0.00
men-dominated power structure	0.00	0.00	0.00	0.00	0.00	0.00
mocking	0.00	0.00	0.00	0.00	0.00	0.00
rejection of feminism	0.00	0.00	0.00	0.00	0.00	0.00
sexual objectification	0.23	0.48	0.31	0.08	0.16	0.10
stereotype generalization	0.00	0.00	0.00	0.00	0.00	0.00
victim blaming	0.02	0.67	0.04	0.11	0.44	0.17
whataboutism	0.12	0.11	0.11	0.03	0.03	0.03
Mistral2						
aggressive and violent attitude	0.47	0.23	0.31	0.08	0.20	0.11
beauty standard expectations	0.00	0.00	0.00	0.00	0.00	0.00
conservative limiting	0.00	0.00	0.00	0.00	0.00	0.00
derogatory and belittling	0.00	0.00	0.00	0.00	0.00	0.00
double standard	0.00	0.00	0.00	0.00	0.00	0.00
gender distinction	0.00	0.00	0.00	0.00	0.00	0.00
men-dominated power structure	0.00	0.00	0.00	0.00	0.00	0.00
mocking	0.00	0.00	0.00	0.00	0.00	0.00
rejection of feminism	0.00	0.00	0.00	0.00	0.00	0.00
sexual objectification	0.40	0.52	0.45	0.21	0.38	0.27
stereotype generalization	0.00	0.00	0.00	0.00	0.00	0.00
victim blaming	0.05	0.33	0.08	0.20	0.31	0.24
whataboutism	0.08	0.02	0.03	0.08	0.05	0.07
LLaMA						
aggressive and violent attitude	0.13	0.06	0.09	0.00	0.00	0.00
beauty standard expectations	0.00	0.00	0.00	0.00	0.00	0.00
conservative limiting	0.00	0.00	0.00	0.00	0.00	0.00
derogatory and belittling	0.00	0.00	0.00	0.00	0.00	0.00
double standard	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

	EN			IT		
Category	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
gender distinction	0.00	0.00	0.00	0.00	0.00	0.00
men-dominated power structure	0.00	0.00	0.00	0.00	0.00	0.00
mocking	0.00	0.00	0.00	0.00	0.00	0.00
rejection of feminism	0.00	0.00	0.00	0.00	0.00	0.00
sexual objectification	0.08	0.05	0.06	0.00	0.00	0.00
stereotype generalization	0.00	0.00	0.00	0.00	0.00	0.00
victim blaming	0.01	0.17	0.01	0.00	0.00	0.00
whataboutism	0.14	0.15	0.15	0.00	0.00	0.00
LLaMA3						
aggressive and violent attitude	0.67	0.11	0.18	0.00	0.00	0.00
beauty standard expectations	0.00	0.00	0.00	0.46	0.34	0.39
conservative limiting	0.00	0.00	0.00	0.38	0.10	0.16
derogatory and belittling	0.00	0.00	0.00	0.27	0.46	0.34
double standard	0.00	0.00	0.00	0.00	0.00	0.00
gender distinction	0.00	0.00	0.00	0.14	0.44	0.22
men-dominated power structure	0.00	0.00	0.00	0.00	0.00	0.00
mocking	0.00	0.00	0.00	0.27	0.03	0.05
rejection of feminism	0.00	0.00	0.00	0.56	0.09	0.15
sexual objectification	0.34	0.39	0.36	0.24	0.44	0.31
stereotype generalization	0.00	0.00	0.00	0.21	0.23	0.22
victim blaming	0.02	0.67	0.05	0.10	0.78	0.18
whataboutism	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-mini						
aggressive and violent attitude	0.63	0.40	0.49	0.11	0.70	0.19
beauty standard expectations	0.00	0.00	0.00	0.47	0.53	0.50
conservative limiting	0.00	0.00	0.00	0.59	0.12	0.19
derogatory and belittling	0.00	0.00	0.00	0.28	0.16	0.20
double standard	0.00	0.00	0.00	0.03	0.56	0.06
gender distinction	0.00	0.00	0.00	0.04	0.44	0.07
men-dominated power structure	0.00	0.00	0.00	0.08	0.13	0.10
mocking	0.54	0.68	0.60	0.32	0.17	0.22
rejection of feminism	0.00	0.00	0.00	0.39	0.59	0.47
sexual objectification	0.35	0.89	0.50	0.13	0.84	0.22
stereotype generalization	0.00	0.00	0.00	0.22	0.83	0.34
victim blaming	0.04	0.83	0.07	0.10	0.72	0.18
whataboutism	1.00	0.04	0.08	1.00	0.05	0.10

Table 8: Detailed classification reports for EN and IT in zero-shot settings.