scientific reports

Check for updates

OPEN

Women who hate men: a comparative analysis across extremist Reddit communities

Erica Coppolillo[1,2]✉

In the present online social landscape, while misogyny is a well-established issue, misandry remains significantly underexplored. In an effort to rectify this discrepancy and better understand the phenomenon of *gendered hate speech*, we analyze four openly declared misogynistic and misandric Reddit communities, examining their characteristics at a *linguistic*, *emotional*, and *structural* level. We investigate whether it is possible to devise substantial and systematic discrepancies among misogynistic and misandric groups when heterogeneous factors are taken into account. Our experimental evaluation shows that no systematic differences can be observed when a double perspective, both male-to-female and female-to-male, is adopted, thus suggesting that gendered hate speech is not exacerbated by the perpetrators' gender, indeed being a common factor of noxious communities.

In contemporary society, social networks and virtual platforms such as Twitter/X, Facebook, LinkedIn, and Reddit serve as vital conduits for interconnecting individuals, facilitating content sharing, and fostering the exchange of ideas and perspectives. While initially viewed through an optimistic lens as vehicles for positive outcomes like community building, remote collaboration, and online activism, extensive literature highlights their unforeseen and adverse impacts. These include phenomena such as radicalization[1], echo chambers[2], discrimination[3], and misinformation spread[4].

Indeed, despite online platforms affording users the benefits of free speech and anonymity, they also constitute fertile ground for the dissemination of hate speech[5,6]. Hate speech can be defined as "an offensive kind of communication mechanism that expresses an ideology of hate using stereotypes"[7]. It generally targets individuals who belong to protected communities and/or minorities, according to features such as ethnicity, religion, physical characteristics, sexual orientation, class, or gender[8,9]. In particular, *gendered hate speech* refers to the form of hate speech perpetrated due to the gender of the target individuals, and generally intended towards women and girls[7,10,11].

In this context, while the global discourse on online *misogyny* is well-established[12–15], the phenomenon of *misandry* remains a relatively underexplored and insufficiently acknowledged facet of the researched digital landscape.

By resorting to different sources[16] (https://en.wikipedia.org/wiki/Misogyny, https://www.britannica.com/topic/misogyny), we can define misogyny as the "hatred or prejudice against women or girls. A form of sexism that has taken shape in multiple forms such as male privilege, patriarchy, [...] discrimination, and sexual objectification". Similarly, misandry denotes the "hatred, dislike, contempt for, prejudice against men or boys." (https://www.dictionary.com/browse/misandry). As also suggested by Wikipedia (https://en.wikipedia.org/wiki/Misandry), despite being both widespread and detrimental phenomena, misandry does not have the same level of institutional and systemic support as misogyny, in nearly all societies.

To grasp an intuition in this regard in the academic context, we performed the query "online misogyny" (resp. "online misandry") on Google Scholar, Scopus, and WebOfScience, and further compared the results. To the query "online misogyny", the former produces $\sim$ 150.000 results, against the $\sim$ 19.600 obtained by searching for "online misandry". As an example, Figure 1 shows the query results fetched on WebOfScience.

As we can see, the number of records referring to misandry are negligible compared to the respective counterparts, with a maximum gap of two orders of magnitude. Similarly, Fig. 2 reports the results obtained when the same queries are performed on Scopus. The time trend shows that (i) the phenomenon of online misogyny is significantly more studied than misandry, and (ii) the latter started being addressed in 2014 only, far 13 years after the first work about online misogyny appeared.

[1]Department of Mathematics and Computer Science, University of Calabria, 87036 Rende, Italy. [2]ICAR-CNR, 87036 Rende, Italy. ✉email: erica.coppolillo@unical.it

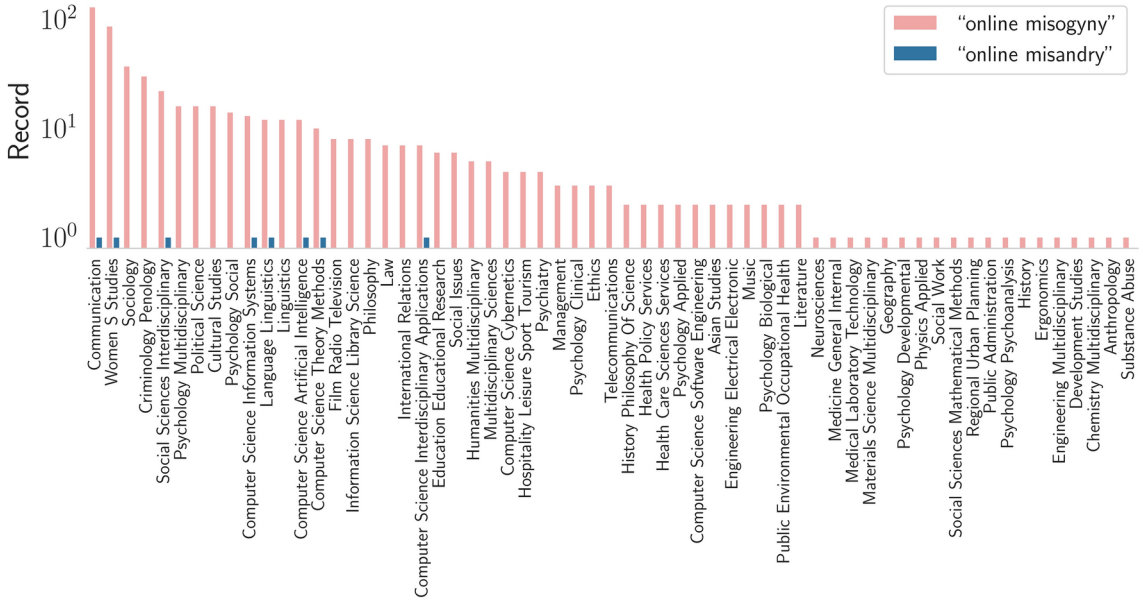


Figure 1. Records distribution over several research categories when the queries "online misogyny" (pink) and "online misandry" (blue) are performed on WebOfScience. The X-axis reports the categories while the Y-axis shows the number of fetched records (in log-scale).

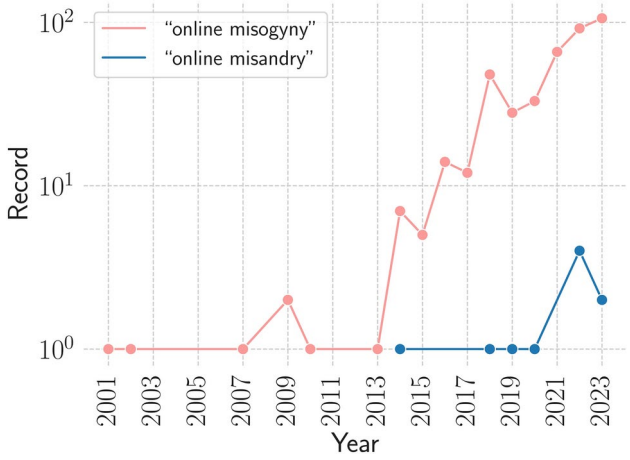


Figure 2. Records distribution over time when the queries "online misogyny" (pink) and "online misandry" (blue) are performed on Scopus. The X-axis reports the publication year while the Y-axis shows the number of fetched records (in log-scale).

In an effort to better understand this phenomenon, our work embarks on a extensive comparative analysis across four extremist communities on Reddit.

Reddit is a social media platform where users can submit content, such as text posts, links, images, and videos, to themed communities known as "subreddits." These subreddits cover a vast range of topics, from news and science to memes and hobbies. Users can upvote or downvote posts and comments, determining their visibility and popularity. Reddit fosters discussion and interaction through comments, allowing users to engage with each other in threads beneath posts. We chose to study this platform in particular, because (i) it is still one of the most active social communities for online discussion, and (ii) differently from others in which the communication takes place as person-to-person only (e.g., Twitter/X), Reddit allows for broader discussions, to which multiple users can participate.

Our main objective consists of addressing the following question: *Can we discern systematic discrepancies among the groups?* In other words, *is gendered hate speech influenced by the perpetrators' gender, or is it a broad-based phenomenon common to toxic communities?* To the best of our knowledge, this is the first attempt to examine gendered hate speech manifestation from both male-to-female and female-to-male perspectives. By shedding light on this crucial aspect of online discourse, we strive to contribute to a more informed and

equitable digital landscape. Our focus extends beyond the academic and research spheres, encompassing broader societal implications. We believe that examining noxious phenomena such as gender discrimination and online hate speech can have a significant impact on multiple aspects of society, uncovering existing stereotypes and misconceptions. A more thorough investigation in this area could have practical implications for policy-making, platform moderation, and the development of interventions that safeguard the well-being of all users, irrespective of gender. For instance, automated detection systems could be enhanced to detect harmful language related to both misogynist and misandrist subcultures. This includes not just direct hate speech, but also subtler forms of sexism and incitement to violence (e.g., perpetration of stereotypes). More specifically, online platforms could integrate moderation systems that are sensitive to the specific language and coded terms used by extremist communities. Further, they should also consider proactive interventions that offer supportive resources (e.g., mental health assistance) to users who frequent misogynist and misandrist related forums, rather than only punitive measures like bans.

The rest of the paper is structured as follows. The Related Work section presents an overview of the current literature focused on gendered hate speech on social media. Further, we formalize the evaluation framework adopted to address our research questions. The Analysis section reports the results of our investigation. In the Data Section, we provide specific information related to the data we analyzed, while discussing the work's limitations in the corresponding section. Finally, the last section concludes the work and devises pointers for further research.

Related work

Social media platforms serve as primary conduits for the propagation of online harassment and gender-based social hate, as underscored by Abarna et al.[12]. Within feminist scholarship, gendered online harassment is perceived as a reflection of broader cultural norms regarding gender roles and the subordinate position of women in society, as evidenced by several studies[14,17]. Jane et al.[13] posit that gendered online hate speech originates from entrenched misogynistic ideologies that reinforce women's perceived inferiority. Similarly, other works[14,15] argue that gender-based hate speech seeks to reaffirm men's perceived natural dominance, particularly in response to perceived threats to their relative social status.

The phenomena of gender-based hate speech and sexual harassment have been perceived as a means of asserting and maintaining traditional gender norms and power differentials, effectively transplanting offline misogyny into the digital realm. Additionally, online harassment poses specific challenges due to its potential anonymity and longevity, allowing a single harmful comment to propagate across platforms indefinitely[18].

Moreover, various groups have been identified to be particularly vulnerable to misogynistic content, including female politicians, journalists, celebrities, influencers, musicians, gamers, YouTubers, and university students[19]. For example, Silva-Paredes and Ibarra Herrera[20] conducted a critical discourse analysis of the abuse directed at a Chilean right-wing politician, while Phipps and Montgomery[21] examined the portrayal of Nancy Pelosi in misogynistic attack ads during Donald Trump's 2020 campaign. Ritchie[22] highlighted how media continues to create harmful representations of female politicians like Hillary Clinton, with significant consequences for political campaigns and democratic processes. Similarly, studies conducted by Wagner[23] and García-Díaz[24] suggest that online harassment is a gendered phenomenon, as women are more aware of and affected by it, with young women particularly vulnerable to sexualized forms of harassment. Further, while Saluja and Thilaka[25] revealed the distinct patterns of harassment female politicians in India face, Fuchs and Schäfer[26] also studied misogynistic hate speech and abuse against female politicians in Japan. Regarding female journalists, Chen et al.[27] revealed that gendered harassment, including sexist comments and threats of sexual violence, is a common experience across various countries, often limiting their engagement with audiences. Koirala[28] found similar patterns in Nepal, where many female journalists either endure harassment or avoid social media. Rego[29] (2018) echoed these findings, focusing on Indian journalists and the prevalence of harassment on social media. In the realm of entertainment, Ghaffari[30] showed how female celebrities on Instagram face pressures to conform to stereotypical gender roles, while Döring and Mohseni[31] found that female YouTube video producers receive more negative comments when addressing feminist topics or displaying their sexuality, though they face less criticism when conforming to traditional roles.

Further, numerous studies on online misogyny focus on the "Manosphere" a network of websites and social media groups that promote misogynistic beliefs. These groups are diverse in their ideologies and levels of violence, often aligning with far-right, homophobic, and racist views[32,33]. Despite their differences, they share the portrayal of feminism as inherently discriminatory and threatening to men[34]. Central to these communities is the concept of a "gynocentric order" and the "red pill" ideology, where members believe they have awakened to an oppressive reality dominated by women. A common theme is the use of the term "misandry", which serves as a tool for community-building and reinforces a belief that feminism is hostile toward men[35]. This narrative is adopted by both extremist misogynist groups and moderate men's rights advocates, who position men as victims of reverse discrimination[34]. Men's far-rights activists, for instance, frame their arguments to construct a narrative of male oppression while denying the existence of gendered violence[36]. Within the Manosphere, ideologies often normalize and legitimize violence against women, trivialize rape accusations, and present men as victims who need to restore patriarchal values[33]; Garcìa-Mingo et al.[37]. For example, Incels ("involuntary celibates") justify their lack of sexual relationships through biological determinism and a belief that feminism and societal norms have victimized them[38]. Studies of incel communities show the use of misogynistic humor, self-deprecating language, and derogatory terms like "femoid" to dehumanize women and justify violence[39,40]. Other subgroups like MGTOW (Men Going Their Own Way) advocate for male separatism, encouraging men to avoid relationships with women and embrace individual empowerment. While less overtly violent, their rhetoric promotes toxic masculinity and reinforces misogynistic views under the guise of rational thinking[41,42]. This body of research highlights the varied and insidious ways in which the Manosphere perpetuates gender-based

hatred and violence, using both extremist and moderate rhetoric to normalize misogyny across different online platforms.

This said, despite concerted efforts have been made to address hate speech targeting women and misogyny in broad sense[43,44], research on online misandry remains relatively scarce, thus not contributing in mitigating existing stereotypes and misconceptions. As an example, Nadim et al.[45] show that more men than women experience online harassment, contrary to prevailing assumptions.

The lack of insights in this respect underscores the importance of conducting comprehensive investigations from both perspectives. Thus, our study seeks to address this gap by analyzing both misogynistic and misandric Reddit communities, aiming to identify and assess potential disparities in linguistic, emotional, and structural features. To the best of our knowledge, this study represents the first attempt to undertake such a dual perspective analysis, making a significant contribution to the understanding of online gender-based harassment. From our analyses, it emerges that the *perspective* under which the misogynistic and misandric communities are compared is crucial for drawing conclusions. For instance, while examining the platform content at a text-level reveals higher toxicity in misogynistic communities, as suggested in[43,44], analyzing the emotion distributions across the subreddits shows skewer levels of hate in misandric sub-populations, as stated in[45]. Our findings therefore indicate that online gender-based hate speech, whether directed at women or men, should be regarded with equal seriousness. It is crucial to recognize that hate speech harms individuals regardless of gender, and therefore, interventions aimed at mitigating such behavior should be gender-neutral. Implementing effective, gender-agnostic solutions can ensure that all forms of hate speech are addressed consistently and fairly, preventing any group from being overlooked. By adopting this approach, platforms and policymakers can foster a more inclusive and respectful digital environment for everyone, promoting equality and safeguarding individuals from gendered abuse.

Formal framework

In this work, we investigate the dynamics of misogynistic and misandric communities on Reddit. Our examination entails a comprehensive analysis of linguistic features, the extraction of emotional nuances embedded within the textual content, and a structural graph-based analysis.

Our goal is addressing the following research questions:

RQ1: Can we devise systematic discrepancy between female-to-male and male-to-female perspectives?

RQ2: Is misogyny over-represented than misandry within extremist Reddit communities?

RQ3: Is gendered hate speech conditioned by the community gender, or indeed consists in a detrimental phenomenon typical of extremist groups?

In other words, we are interested in assessing if and at which extent misogynistic communities differ from misandric ones, either in terms of linguistic, emotional or structural features.

We conducted our analysis on the following four Reddit extremist communities:

- Feminism: *r/Feminism* is a feminist political subreddit discussing women's issues. It has 277,000 users and more than 50% of the posts have been assessed to exhibit a predominantly negative sentiment[46].
- GenderCritical: The subreddit *r/GenderCritical* had 64,400 users and self-described as "reddit's most active feminist community" for "women-centred, radical feminists" to discuss "gender from a gender-critical perspective". In 2020, the subreddit was banned for violating Reddit's rule against promoting hate and transphobia.
- Incels: *r/Incels* was a forum wherein members discussed their lack of partnering success. Many members adhered to the "black pill" ideology, which espoused despondency often coupled with misogynistic views. The subreddit was banned in 2017, and at that time it counted 40,000 subscribers.
- MensRights: Created in 2008, the "antifeminist" subreddit *r/MensRights* has over 300,000 members as of April 2021. It has been recognized as one of "the most striking features of the new antifeminist politics"[47].

GenderCritical, Incels and Mensrights are reported as "Controversial Reddit communities" on Wikipedia (https://en.wikipedia.org/wiki/Controversial_Reddit_communities).

Based on the properties of the aforesaid communities, from now on we will refer to "GenderCritical" and "Feminism" as the *misandric* communities, and to "Incels" and "Mensrights" as the *misogynistic* ones.

Our choice of misandric representatives is further justified by the taxonomy of Online Women's Ideological Spaces (OWIS) presented by Balci et. al.[48]. Based on their analysis, GenderCritical is reported into the "Gender-Critical Feminism" group, described as "a more radical way of feminism that sees anyone with a penis as automatically an oppressor"; while Feminism is categorized within "Mainstream Feminism", a more liberal and mainstream ideology. In the latter, a wide range of topics have been observed, including toxic masculinity, patriarchy and women disadvantages due to toxic-masculine mindsets.

Analysis

Since we are interested in misogynistic (resp. misandric) content that targets women (resp. men), we performed a pre-processing step on the data by (*i*) keeping the posts and comments from misandric subreddits that contain the word "boy", "man", "men", "boyfriend", and/or "husband", and (*ii*) keeping the posts and comments from misogynistic subreddits that contain the word "girl", "woman", "women", "girlfriend", and/or "wife".

This is because, our objective is investigating the linguistic and emotional dynamics perpetrated by misandric women towards men (resp. misogynistic men towards women) only, thus filtering out potential other discussion topics. We further discarded textual content shared by users who had been removed or deleted from the Reddit platform, due to the lack of information on such users. Some data statistics after the cleaning step are reported in Table 1.

Subreddit	#Authors	#Commenters	#Posts	#Comments	#Users	#Texts
Feminism	5795	30461	6735	84648	35390	91383
GenderCritical	1900	17361	3832	254901	17878	258733
Incels	1235	14191	3396	125489	14547	128885
Mensrights	11638	116926	26687	1132521	120978	1159208

Table 1. Data statistics after the cleaning step. "Authors" refer to the users who created a post within the subreddit, while "Commenters" refer to the users who wrote at least a comment.

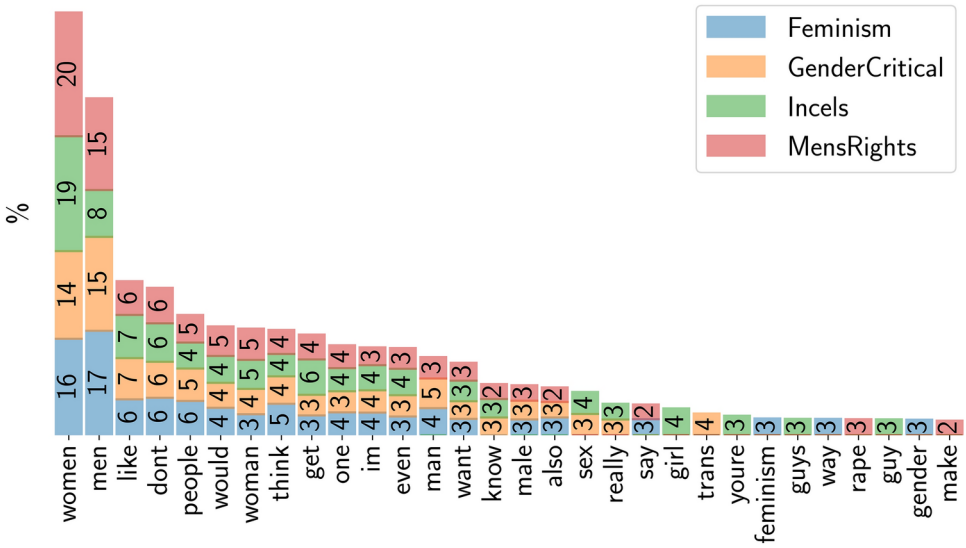


Figure 3. Relative frequency of the words obtained by aggregating the 20-most common words of each community. The X-axis reports the obtained words, while the slots in each bar report the relative frequency (in percentage) of the word in the corresponding community.

Linguistic discrepancy

We first evaluate the differences between the misogynistic and misandric communities at a word-level for estimating linguistic discrepancy.

Setting. We begin by performing an initial cleaning of the text, removing elements like punctuation, digits, and stopwords to eliminate potential noise. For this preprocessing step and subsequent analysis, we use the widely adopted Python library, "nltk" (https://www.nltk.org/), which is commonly employed in natural language processing tasks.

Results. We retrieve the 20-most common words from each subreddit, and compare their relative frequencies. Figure 3 reports the result. The X-axis represents the words obtained by aggregating the 20-most common of each community, while the slot in each bar reports the relative frequency of the word in the corresponding community. We highlight three main observations:

- Only a few words are representative of their community, i.e., are the most frequent exclusively in a specific subreddit (e.g., "rape" in Mensirghts, "trans" for Gendercritical, and "feminism" for Feminism).
- The majority of the words occur with similarly high relative frequency in all the subreddits, with no distinction between misogynistic and misandric.
- The words "women" and "men" appears more often in the misogynistic and misandric subreddits, respectively (as an expected consequence of the filtering step), with relative frequencies of 19% and 20% on Incels and Mensrights for "women", and of 17% and 15% on Feminism and Gendercritical for "men". Notably, these words occur almost likewise frequently in the counterpart communities: in particular, the word "women" occurs on Feminism and Gendercritical with a relative frequency of 16% and 14%, respectively; while the word "men", occurs on Incels and Mensrights in the 8% and 15% of the cases, respectively.

These considerations suggest that no sharp linguistic differences can be identified among misogynistic and misandric subreddit communities at a word level.

Toxicity

We conduct a different textual investigation over the posts and comments of each subreddit, in order to estimate the degree of content *toxicity*.

Setting: For the task, we adopted a version of RoBERTa, a transformer-based text classifier, which has been fine-tuned for hate speech detection[49]. As reported by the authors, the training dataset comprised 41,255 entries, of which 18,993 have been manually annotated as "not hate" and 22,262 as "hate". The posts tagged as "hate" have been in turn divided into sub-categories, such as "Animosity", "Dehumanization", "Derogation", "Support", "Threatening", and "Other". On the test set, the fine-tuned model achieves a Macro F1-score of 75.97 with a standard deviation of 0.96 over 5 training rounds, thus showing high classification reliability. For our purpose, we feed the model with a textual content t and retrieve its toxicity score $s(t)$, spanning in the range [0, 1].

Results: Figure 4 shows the distributions of the computed toxicity scores, one relating to each subreddit. The X-axis reports the score s, while the Y-axis represents the density. As we can see, all the analyzed communities present a bimodal distribution mostly skewed on 0, with another less prominent peak on 1. This suggests the following considerations:

- The majority of the content is estimated as *non-toxic* (0 score), with no distinction on the community.
- The two subreddits that present the highest peak on 1 are the misogynistic ones, Incels being the most toxic.

The reported results indicate that the misogynistic communities indeed convey a higher toxicity degree than misandric ones. Interestingly, however, all the distributions exhibit a similar shape besides the borders, i.e., posts and comments tend to be either non-toxic or extremely toxic.

Prevalent emotions

Up to this point, we analyzed lexical features for assessing discrepancies among the communities, both in terms of lexical properties or language toxicity. Next, we focus our attention on the conveyed emotions that emerge from the content of posts and comments. Our objective is to investigate how emotions span within different subreddits and within each one.

We are especially interested in analyzing the *negative* emotions expressed in the communities. In particular, we restrict our attention on: *sadness*, *anger*, *fear* and *hate*. We denote the aforesaid emotions set as $\mathscr{F}$. More formally, for each textual content t, either post or comment, we provide a set of labels $f(t) \in \mathscr{F}$.

Setting: To accomplish the labelling task $f(t)$, we adopt two different techniques. First, we exploit Google-T5[50], a Text-to-Text Transfer Transformer (T5) fine-tuned for emotion recognition downstream task (https://huggingface.co/mrm8488/t5-base-finetuned-emotion). The model has been trained on the "Emotion Dataset"[51], in order to recognize the following emotions: *sadness*, *joy*, *love*, *anger*, *fear*, and *surprise*. As reported in the reference card, the model achieves great classification performances, with a F1-score of 0.93, 0.89, 0.95, 0.85, 0.97, 0.80 for each of the aforesaid emotions, respectively. Secondly, we adopt Empath[52], a language processing tool that leverages neural embeddings to connect words and phrases by analyzing a large corpus of modern fiction text. To assess the accuracy of the tool, the authors built a dataset which consists of a mixed textual corpus comprising more than 2 million words ranging over 4500 documents. They further compute Pearson correlations between the results produced by Empath with the ones obtained with Linguistic Inquiry and Word Count (LIWC), which has been extensively validated in the literature[53] and hence chosen as reference benchmark. The analysis showed that Empath shares overall average Pearson correlations of 0.90 with LIWC, thus proving reliability. The tool spans over 200 pre-validated categories. For our purposes, we restrict our attention to the emotions in $\mathscr{F}$.

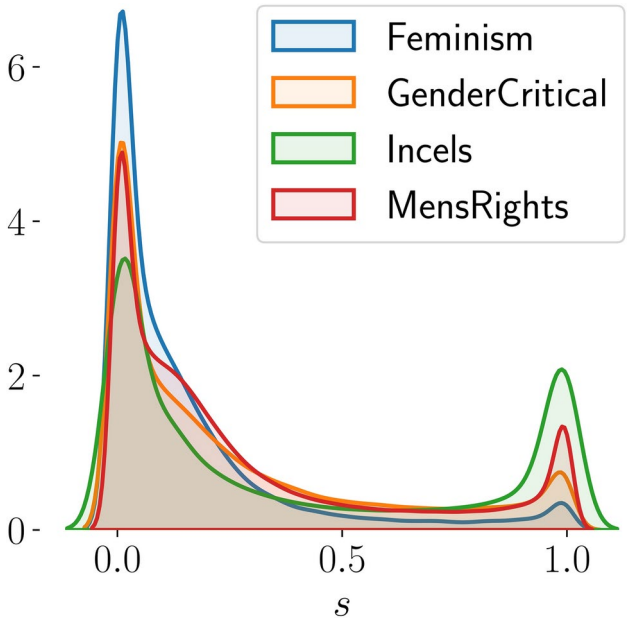


Figure 4. Distribution of the content toxicity of each subreddit. The X-axis represents the toxicity score, while the Y-axis represents the distribution density.

We proceed to label the posts and comments retrieved from each subreddit by adopting the following protocol: first, we check if Empath recognizes in the text any of the emotions in $\mathscr{F}$; if not, the text is discarded; otherwise, we consider as label(s) the union of the emotions detected by both Empath and Google-T5, eventually considering the intersection of the result with $\mathscr{F}$.

This experimental choice is justified by the fact that Google-T5 always produces in output one of the emotions above-mentioned, thus leading to potential false positives (in other words, a text could express none of the emotions in $\mathscr{F}$, but one would be produced anyway). Empath, on the contrary, is able to detect a wider spectrum of emotions, thus producing a 0 score for the emotions of interest, being less prone to false positives. By exploiting the tools jointly, we indeed (i) ensure that the analyzed content actually reflects at least one of the emotions we are interested in, and (ii) capture the potentially complementary outcomes of both algorithms.

Results: The outcome of the labelling procedure is depicted in Figure 5a. As we can see, Feminism and Mensrights present the most similar distributions, skewed on *hate* and *anger*, while GenderCritical slightly sloping towards *fear* and Incels, more evidently, towards *sadness*. Notably, all the communities present a consistent peak towards *hate*. Hence, no significant discrepancy between misogynistic and misandric groups appears. We point out that the results are not contradictory with respect to the analysis reported in the previous section. Indeed, as just described, we here discard any content whose detected emotions do not fall within $\mathscr{F}$. In other words, the portion of texts considered is a subset of the content used to compute the toxicity distribution reported in Fig. 4. We further argue that the content analysis can be biased towards prolific users, who can shift the distributions without presenting the actual subreddit population. Inspired by this observation, we reframe the investigation by considering the emotions at a user-level. In more detail, let u be a user and let $\mathscr{T}^u = \{t_1^u, \ldots, t_n^u\}$ be the set of content (post/comment) associated with u. We can define $\mathscr{F}_j^u = \{t_i^u | j \in f(t_i^u)\}$, as the set of texts produced by u and tagged with emotion j, according to the aforesaid labelling procedure.

Given the natural order $|\mathscr{F}_1^u| \leq \ldots \leq |\mathscr{F}_m^u|$ induced over the sets, we define $f(u) = \{m\}$. In other words, the label $f(u)$ is the emotion occurring in the majority of the texts produced by u.

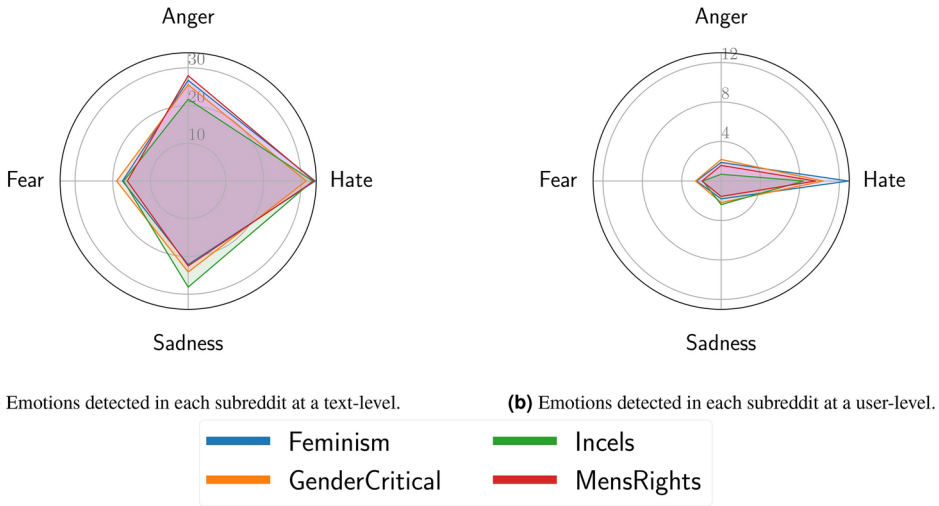
The emotion distributions obtained by the user-level tagging are depicted in Fig. 5b. Despite no significant differences can be observed regarding GenderCritical and Incels (which remain mainly skewed toward *fear* and *sadness*, respectively), the user-level perspective shifts the distributions peaks of Feminism and Mensrights: while at a text-level, the two communities do not show a visible gap, here we can see that Feminism consistently overcomes the others in terms of hate, followed by GenderCritical and Mensrights.

We can hence conclude that:

- Different results can be devised when considering content-level and user-level emotions.
- When adopting a content-level perspective, all the communities show the greatest peak towards *hate*, followed by *anger*. Feminism and Mensrights present the most similar distributions, while Incels and GenderCritical slightly sloping towards *sadness* and *anger*, respectively. No significant differences among the misogynistic and misandric communities can be devised.
- Conversely, when a user-level perspective is taken into account, distributions drastically change, magnifying the *hate* peak of Feminism, which significantly overcomes the other communities. Also GenderCritical, despite maintaining an inclination toward *fear*, skews on *anger* and *hate* as well. Under this optic, indeed, misandric communities express more negative sentiments than misogynistic ones.

Graph-based analysis

We finally conduct a comparison among the subreddits based on the underlying graph structure. For each community, we build the interaction graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, such that $\mathscr{V}$ is the set of users, and $(u, v) \in \mathscr{E}$ if user



(a) Emotions detected in each subreddit at a text-level.

(b) Emotions detected in each subreddit at a user-level.

Figure 5. Prevalent emotions detected in each subreddit either at text-level (**a**) and user-level (**b**). Circumferences represent different percentages.

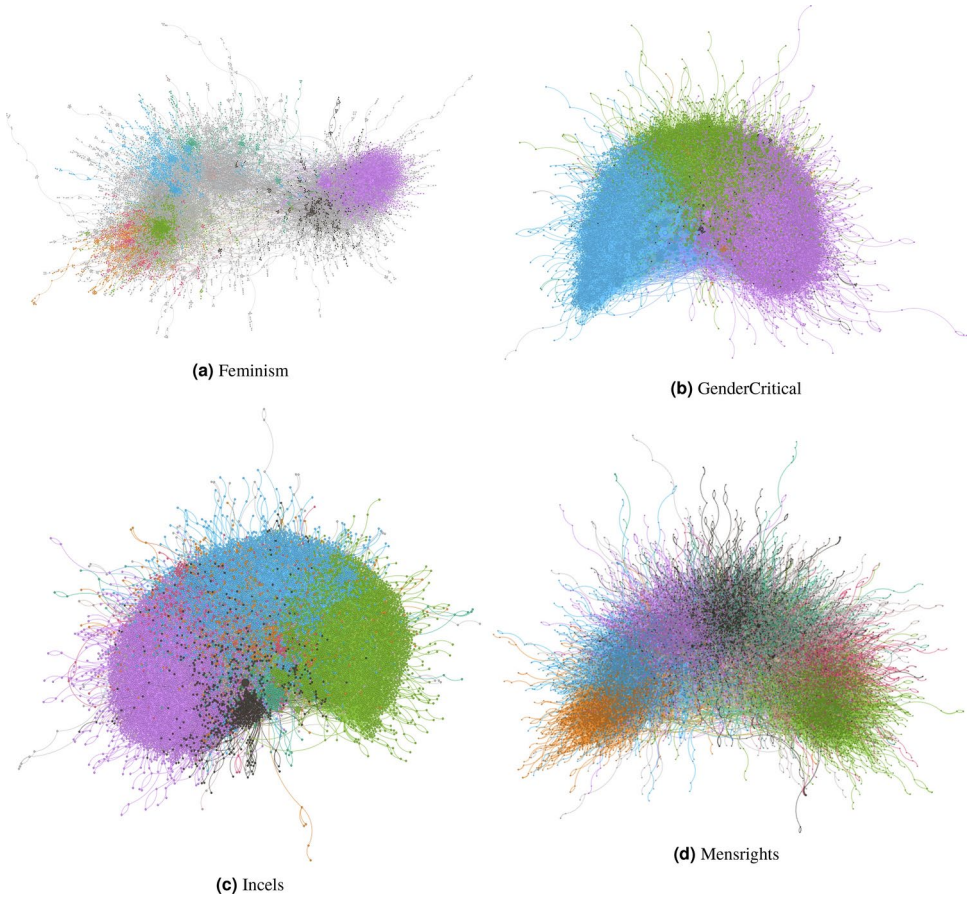


Figure 6. Communities interaction graphs coloured by modularity.

Subreddit	$\lvert\mathscr{V}\rvert$	$\lvert\mathscr{E}\rvert$	Degree	Path Length	Diameter	Modularity
Feminism	19,367	27,309	1.41	6.77	25	0.82
GenderCritical	14,007	90,088	6.43	3.87	10	0.40
Incels	10,707	41,476	3.87	4.0	11	0.45
MensRights	48,029	100,000	2.08	5.18	18	0.57

Table 2. Statistics of the community graphs in terms of number of nodes, number of edges, degree, average path length, diameter and modularity.

u replied to a post or a comment shared by user v. Note that $\mathscr{V}$ is a subset of the total number of users since an interaction is considered (i.e., an edge is added) only if both the pieces of content (either post and comment or comment and comment) contain the filtering words used in the pre-processing step. In other words, an interaction is considered only if it focuses on men in the misandric communities, and women in the misogynistic ones, respectively.

Figure 6 depicts the interaction graphs of the communities, with each community highlighted using color to represent modularity. The modularity values, which indicate the strength of division between the communities, were computed using the Blondel method[54] (also known as the Louvain algorithm). This method is widely recognized for its efficiency in detecting community structures within large networks by optimizing the modularity score. Further, Table 2 reports some graphs statistics (number of nodes, number of edges, average degree, average shortest path length, diameter, and modularity). For Mensrights, we consider a random sample of 100000 edges to ensure computational feasibility in the calculus of modularity. Figure 7 depicts the degree distributions of the considered communities graphs[55], which follow a power-law pattern, a common characteristic observed in social networks[56].

Our objective is indeed to estimate if significant differences can be devised between the misogynistic and misandric sub-reddits in terms of the structure of their interaction graph. Surprisingly, GenderCritical and Incels present similar structural characteristics, both in terms of average path length (3.87 and 4.0, respectively), diameter (10 and 11, respectively) and modularity (0.4 and 0.45, respectively).

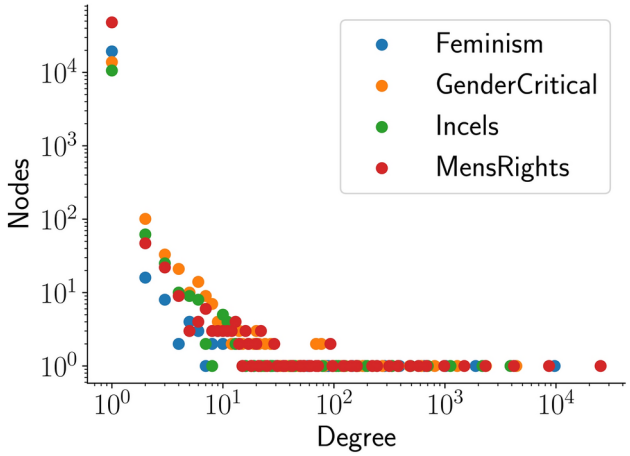


Figure 7. Degree distribution of the communities graphs. The X-axis reports the degree, while the Y-axis shows the number of nodes. Both axes are in log-scale.

Feminism, indeed, shows the highest modularity (0.82), the longest average path (6.77) and the highest diameter, despite having the lowest average degree (1.41). Notably, its properties do not substantially differ from Mensrights: indeed, the latter presents the second-highest modularity, average path length and diameter (resp., 0.57, 5.18 and 18), as well as the second-lowest average degree (2.08).

This suggests that Feminism interaction graph shares more structural features with Mensrights than with the other misandric community (i.e., GenderCritical), and similarly, Incels is significantly more similar to GenderCritical than to the other misogynistic group (i.e., Mensrights). We can hence conclude that no heterogeneous structural features can be devised in terms of interaction graphs across misogynistic and misandric communities.

Data

The data used in this study originates from the PushShift archive[57], a publicly accessible dataset that has been referenced in over a hundred peer-reviewed publications. It contains comments and posts collected between January 1, 2016, and December 31, 2022, sourced from the 20,000 most popular subreddits. To safeguard user anonymity, no personally identifiable information was utilized in the selection or review of users, and all analyses were conducted and reported at an aggregate level. As the study involved no direct human interaction, it did not require Institutional Review Board (IRB) approval.

Limitations

The first limitation of this study stems from the reliance on open-source platform data, which is inherently prone to noise.These datasets often contain inconsistencies, irrelevant information, or errors that can affect the quality of the analysis. To address this, as detailed in the Analysis Section, we applied a pre-processing pipeline to clean the textual data by removing punctuation, digits, and stop words. While this step significantly reduces noise, it cannot guarantee perfectly clean data. The dynamic and unstructured nature of open-source data makes achieving complete purification impractical in real-world scenarios. Additionally, the outcomes presented in this study are highly dependent on the performance of the classifiers employed to assess content toxicity (subsection "Toxicity") and emotions (subsection "Prevalent emotions"). Although the reported metrics indicate that these models perform well, as discussed in the respective sections, it is essential to recognize their inherent limitations. These classifiers, while advanced, are not flawless and should not be regarded as definitive arbiters of content interpretation. Misclassifications, therefore, are an unavoidable reality. While their frequency is low and their impact on the final results is expected to be minimal, such errors nonetheless introduce a degree of uncertainty that could slightly affect the overall accuracy and reliability of the findings. As a final note, it is important to emphasize that the findings presented in this paper are specific to the examined extremist communities on Reddit and should not be generalized to other social platforms. The results are inherently context-dependent and may differ if the methodology is applied to other networks, such as Facebook or Twitter/X, or to different groups of interest. This highlights the importance of tailoring analyses to the unique characteristics and dynamics of each platform and community under study.

Conclusions and future work

In this work, we addressed the detrimental phenomenon of gendered hate speech online, by conducting extensive analyses across four extremist Reddit communities, two of which openly declared misogynistic and misandric, respectively. We conducted our analysis at a linguistic, emotional and structural level: first, we evaluated the most common words and the content toxicity adopted and shared within each community; second, we compared the prevalent emotions both at a text- and user-level, with respect to negative feelings such as *hate*, *anger*, *fear* and *sadness*; lastly, we constructed the interaction graph of each community, studying their structural properties.

The performed analyses reveal that no systematic differences can be devised across the misogynistic and misandric communities. This suggests that, in addressing the phenomenon of online gendered hate speech, both male-to-female and female-to-male perspectives should be taken into account, thus recognizing equal importance to both *misandry* and *misogyny*.

The analysis devised in this paper can be further strengthened by additional research. We identify, among the others, (i) investigating the level of disagreement/homophily within the discussion hubs by performing analysis cascades; (ii) assessing if fake users (e.g., bots) contribute to negative contamination and hate speech spread; and (iii) identifying potential hierarchies, i.e., radicalized/polarized subgroups within the same community, further comparing their impact within the misogynistic vs misandric groups.

Data availability.

The datasets generated and/or analysed during the current study are not publicly available due to privacy issue but are available from the corresponding author on reasonable request. The code is available at the following github: https://github.com/EricaCoppolillo/WomenWhoHateMen.

Received: 4 July 2024; Accepted: 27 November 2024
Published online: 22 April 2025

References

1. Phadke, S., Samory, M. & Mitra, T. Pathways through conspiracy: The evolution of conspiracy radicalization through engagement in online conspiracy discussions. *Proc. Int. AAAI Conf. Web Soc. Media* **16**, 770–781 (2022).
2. Pariser, E. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think* (Penguin Books, 2012).
3. Tao, X. & Fisher, C. B. Exposure to social media racial discrimination and mental health among adolescents of color. *J. Youth Adolesc.* **51**, 30–44 (2022).
4. Aimeur, E., Amri, S. & Brassard, G. Fake news, disinformation and misinformation in social media: A review. *Soc. Netw. Anal. Min.* **13** (2023).
5. Guiora, A. N. & Park, E. A. Hate speech on social media. *Philosophia* **45**, 957–971 (2017). https://api.semanticscholar.org/CorpusID:148707841.
6. Mathew, B., Dutt, R., Goyal, P. & Mukherjee, A. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19. 173–182 (Association for Computing Machinery, 2019).
7. Chetty, N. & Alathur, S. Hate speech review in the context of online social networks. *Aggress. Violent Behav.* (2018).
8. Elsherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y. & Belding-Royer, E. M. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *International Conference on Web and Social Media* (2018). https://api.semanticscholar.org/CorpusID:4809781.
9. Silva, L. A., Mondal, M., Correa, D., Benevenuto, F. & Weber, I. Analyzing the targets of hate in online social media. In *International Conference on Web and Social Media* (2016). https://api.semanticscholar.org/CorpusID:10634337.
10. Ging, D. & Siapera, E. *Gender Hate Online Understanding the New Anti-Feminism: Understanding the New Anti-Feminism* (Palgrave Macmillan Cham, 2019).
11. Saresma, T., Karkulehto, S. & Varis, P. Gendered violence online: Hate speech as an intersection of misogyny and racism. *Violence Gender Affect* (2020). https://api.semanticscholar.org/CorpusID:234341127.
12. Abarna, S., Sheeba, J., Jayasrilakshmi, S. & Devaneyan, S. P. Identification of cyber harassment and intention of target users on social media platforms. *Eng. Appl. Artif. Intell.* **115**, 105283 (2022). https://www.sciencedirect.com/science/article/pii/S0952197622003359.
13. Jane, E. A. 'back to the kitchen, cunt': Speaking the unspeakable about online misogyny. *Continuum* **28**, 558–570 (2014).
14. Perry, B. *In the Name of Hate: Understanding Hate Crimes* (Routledge, 2001).
15. Berdahl, J. L. Harassment based on sex: Protecting social status in the context of gender hierarchy. *Acad. Manag. Rev.* **32**, 641–658 (2007).
16. Srivastava, K., Chaudhury, S., Bhat, P. & Sahu, S. Misogyny, feminism, and sexual harassment. *Indus. Psychiatry J.* **26**, 111 (2017).
17. Jane, E. A. *Misogyny Online: A Short (and Brutish) History* (Sage, 2016).
18. Mathew, B. et al. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media* (2018). https://api.semanticscholar.org/CorpusID:52002120.
19. Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A. & Tontodimamma, A. How do we study misogyny in the digital age? a systematic literature review using a computational linguistic approach. *Hum. Soc. Sci. Commun.* **11**, 1–15 (2024). https://api.semanticscholar.org/CorpusID:268857783.
20. Silva-Paredes, D. & Herrera, D. I. Resisting anti-democratic values with misogynistic abuse against a Chilean right-wing politician on twitter: The #camilapeluche incident. *Discourse Commun.* **16**, 426–444 (2022). https://api.semanticscholar.org/CorpusID:250397518.
21. Phipps, E. & Montgomery, F. "Only you can prevent this nightmare, America": Nancy pelosi as the monstrous-feminine in Donald Trump's Youtube Attacks. *Women's Stud. Commun.* **45**, 316–337 (2022). https://api.semanticscholar.org/CorpusID:251171869.
22. Ritchie, J. Creating a monster. *Feminist Med. Stud.* **13**, 102–119 (2013). https://api.semanticscholar.org/CorpusID:142886430.
23. Wagner, A. Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. *Feminist Med. Stud.* **22**, 32 – 47 (2020). https://api.semanticscholar.org/CorpusID:216247718.
24. García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R. & Valencia-García, R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.* **114**, 506–518 (2021). https://www.sciencedirect.com/science/article/pii/S0167739X20301928.
25. Saluja, N. Women leaders and digital communication: Gender stereotyping of female politicians on Twitter (2021). https://api.semanticscholar.org/CorpusID:259106486.
26. Fuchs, T. H. & Schäfer, F. Normalizing misogyny: Hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum* **33**, 553 – 579 (2020). https://api.semanticscholar.org/CorpusID:213983501.
27. Chen, G. M. et al. 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism* **21**, 877 – 895 (2020). https://api.semanticscholar.org/CorpusID:149786056.
28. Koirala, S. Female journalists' experience of online harassment: A case study of nepal. *Med. Commun.* **8**, 47–56 (2020). https://api.semanticscholar.org/CorpusID:213312934.
29. Rego, R. Changing forms and platforms of misogyny: Sexual harassment of women journalists on Twitter. *Med. Watch* **9**, 472 – 485 (2018). https://api.semanticscholar.org/CorpusID:199569056.
30. Ghaffari, S. Discourses of celebrities on Instagram: Digital femininity, self-representation and hate speech. *Crit. Discourse Stud.* **19**, 161–178 (2020). https://api.semanticscholar.org/CorpusID:229502391.

31. Döring, N. & Mohseni, M. R. Male dominance and sexism on Youtube: Results of three content analyses. *Feminist Media Stud.* **19**, 512–524 (2019).
32. Cervi, L. & Tejedor, S. Borders as the ultimate (de)fence of identity: An ontological security approach to exclusionary populism in Italy and Spain. *KOME* (2022). https://api.semanticscholar.org/CorpusID:246342518.
33. Dickel, V. & Evolvi, G. "Victims of feminism": Exploring networked misogyny and #metoo in the manosphere. *Feminist Media Stud.* **23**, 1392–1408 (2022). https://api.semanticscholar.org/CorpusID:246582242.
34. Farci, M. & Righetti, N. Italian Men's Rights Activism and Online Backlash Against Feminism (2019). https://api.semanticscholar.org/CorpusID:216664707.
35. Marwick, A. E. & Caplan, R. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Stud.* **18**, 543 – 559 (2018). https://api.semanticscholar.org/CorpusID:149246142.
36. Carian, E. K. "We're all in this together": Leveraging a personal action frame in two men's rights forums. *Mobil. Int. Q.* (2022). https://api.semanticscholar.org/CorpusID:247847703.
37. García-Mingo, E., Diaz-Fernandez, S. & Tomás-Forte, S. (Re)configurando el imaginario sobre la violencia sexual desde el antifeminismo: El trabajo ideológico de la manosfera española. *Polít. Soc.* **59**, e80369 (2022).
38. Lindsay, A. Swallowing the black pill: Involuntary celibates' (incels) anti feminism within digital society. *Int. J. Crime Justice Soc. Democracy* (2022). https://api.semanticscholar.org/CorpusID:247174289.
39. di Carlo, G. S. An analysis of self-other representations in the incelosphere: Between online misogyny and self-contempt. *Discourse Soc.* **34**, 3 – 21 (2022). https://api.semanticscholar.org/CorpusID:249304267.
40. Chang, W. The monstrous-feminine in the incel imagination: Investigating the representation of women as "femoids" on /r/braincels. *Feminist Media Stud.* **22**, 254 – 270 (2020). https://api.semanticscholar.org/CorpusID:225421230.
41. Jones, C., Trott, V. A. & Wright, S. Sluts and soyboys: Mgtow and the production of misogynistic online harassment. *New Media Soc.* **22**, 1903 – 1921 (2019). https://api.semanticscholar.org/CorpusID:210530415.
42. Wright, S., Trott, V. A. & Jones, C. 'The pussy ain't worth it, bro': Assessing the discourse and structure of mgtow. *Inf. Commun. Soc.* **23**, 908 – 925 (2020). https://api.semanticscholar.org/CorpusID:219023052.
43. Saha, P., Mathew, B., Goyal, P. & Mukherjee, A. Hateminers : Detecting hate speech against women. arXiv: abs/1812.06700 (2018). https://api.semanticscholar.org/CorpusID:56459439.
44. Richardson-Self, L. Woman-hating: On misogyny, sexism, and hate speech. *Hypatia* **33**, 256–272 (2018).
45. Nadim, M. & Fladmoe, A. Silencing women? Gender and online harassment. *Soc. Sci. Comput. Rev.* **39**, 245–258 (2021).
46. Dilkes, J. Rule 1: Remember the human. a socio-cognitive discourse study of a reddit forum banned for promoting hate based on identity. *Discourse Soc.* **35**, 48 – 65 (2023). https://api.semanticscholar.org/CorpusID:260910477.
47. Ging, D. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men Masculinities* (2019).
48. Balci, U. et al. Beyond fish and bicycles: Exploring the varieties of online women's ideological spaces. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23. 43–54 (Association for Computing Machinery, 2023). https://doi.org/10.1145/3578503.3583618.
49. Vidgen, B., Thrush, T., Waseem, Z. & Kiela, D. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL* (2021).
50. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* (2020).
51. Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J. & Chen, Y.-S. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Riloff, E., Chiang, D., Hockenmaier, J. & Tsujii, J. eds.) (Association for Computational Linguistics, 2018). https://aclanthology.org/D18-1404.
52. Fast, E., Chen, B. & Bernstein, M. S. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16 (Association for Computing Machinery, 2016).
53. Pennebaker, J., Francis, M. & Booth, R. *Linguistic Inquiry and Word Count (LIWC)* (1999).
54. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).
55. Golbeck, J. Chapter 3—Network structure and measures. In *Analyzing the Social Web* (Golbeck, J. ed.). 25–44 (Morgan Kaufmann, 2013). https://www.sciencedirect.com/science/article/pii/B9780124055315000031.
56. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).
57. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. The pushshift reddit dataset. https://arxiv.org/abs/2001.08435. (2008).

Acknowledgements

This work has been partially funded by MUR on D.M. 352/2022, PNRR Ricerca, CUP H23C22000440007. It was also partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU. We sincerely thank Ece Calikus from the Royal Institute of Technology, Stockholm, for providing the Reddit data, and Giuseppe Manco, from ICAR-CNR, Rende, for providing valuable suggestions in writing the paper.

Author contributions

Concept, design, writing, implementation and analysis of the experiments described in the manuscript have been done by Erica Coppolillo.

Declarations

Competing interests.

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.