Online Misogyny Annotation Framework

May 2020

Table of Contents

Front matter

General advice

- When you're unsure about an annotation go back to the two basic questions: 1) Is it misogynistic? 2) It is abusive? Unless it's person-directed abuse we're only interested in content that is misogynistic. If a piece of content doesn't fit those criteria you should annotate it as None. If you're still unsure you can flag it in the comments.
- Remember that content can fall into multiple categories .
- Keep the codebook open at all times and always check. This is not a memory test.
- Be prepared for the expert facilitated meetings and reflect on edge cases/tricky results.
- At the meetings, please share any feedback, in particular actionable insights. This is a collaborative project and the more you feedback, the more we can advance the research.
- See the guidelines on how to have a good Expert facilitated meeting.
- If you read something which is non-English then do not translate it but just make the annotation based on any English text in the entry. If you can't make an annotation because there is no English text then just move on to the next entry.
    - One exception is non-English terms which are used in English; such as 'uber' for 'very', 'wunderbar' for wonderful and 'untermensch' for 'subhuman'.

- If there are links – do NOT click on them. This is because (1) many people in these Subreddits don't click on the links! And (2) it introduces lots of 'noise' for our subsequent analyses. However, you can make reasonable inferences about the content of links based on how they are discussed in the thread. You can also read the link title, which often contains the key information.
- Do Google. If you don't know something or a term seems unfamiliar, please find out what it is. This is particularly important for internet-specific slang. Sources such as Urban Dictionary, Wikipedia, and RationalWiki are particularly helpful.

Reviewing images

Please review all images embedded in the threads. They contain vital information and often set the tone for a conversation, especially when they are in the original post. Annotate images based on how they are deployed by the user. This will involve careful examination: images and memes can be used to express abuse in a range of ways. Often humorous content will be identity-directed abuse in the 'animosity' category, but this is not always the case. Because you cannot highlight an image, *you must write out the relevant bits of text*.

One special case to look out for is when images contain abusive content which has been shared by a user who then distances/critiques that content.

- For instance, someone might share a screenshot of a conversation between two users in which they engage in person-directed abuse. This *by itself* should not be marked up as abusive. The person we are annotating for is the poster in the conversation thread. Sharing interpersonal abusive content produced by others is not necessarily abuse in-itself.
- However, the author might then comment on the image, and by doing so express abuse themselves.
- Often, people share prejudicial memes and images. These should be marked up as abusive.
- However, in some cases authors will share screenshots of other people engaging in abuse, such as a conversation from somewhere else on Reddit. This requires careful reading: if the author explicitly distance themselves from the content of the post then it would not count as abuse – but if they do not distance themselves (which includes not commenting anything at all) then you should mark it up as abuse.
- The key aspect to annotate is how *the author of the entry* uses the image. It is their content you are annotating, and that is what you need to make a decision about.
- If you are unsure about an image, flag it in the comments.

Data quality

If you encounter any data which you think is *wrong*, as in it has been incorrectly collected, wrangled, cleaned or presented, please flag it and notify the project lead immediately. Skip to the next conversation and continuing annotating. If you then encounter more data that you

think is wrong, please stop working as there may be a fundamental problem with the data you have been given.

Annotation codebook

*** Please note that the following section will contain examples of abusive content, using uncensored language. ***

0. None of the other categories

- Content which does not fall into any of the other categories outlined here; it usually is entirely unrelated to abuse or women.
- For instance, ‘Had a great ice cream earlier’, ‘Why is the sky blue?’, ‘Fossil fuel crisis won’t solve itself!’.
- If you think an entry includes misogynistic abuse, but doesn’t fall into one of the categories described, make a note in the comments and we will discuss it during group facilitation
- If you mark an entry as falling within this category then you don’t need to do anything more and can move on to the next entry.

1. Misogynistic pejoratives

- A pejorative is a derogatory term which expresses negative connotations about someone. Misogynistic pejoratives are terms that are often used to disparage women This includes terms which are explicitly insulting and derogatory, such as ‘slut’, ‘whore’, ‘cow’, as well as terms which are not as explicit but nonetheless still implicitly express negativity/animosity against women, such as ‘Stacy’ or ‘Becky’. If there are any terms you are unfamiliar with or unsure of we encourage you to use a search engine, search sites like urban dictionary, etc. to get more context
- To be included in this category the term must be a noun used to negatively refer to a person. For example “whore” should be annotated as a misogynistic pejorative but “whoring” should not as it is a verb used to describe an action, not to directly describe a person. Similarly “pussy” is a misogynistic pejorative but “pussy pass” is not because it is not used to describe a person but a concept or object. However, entries that use such terms (e.g. “whoring” or “pussy pass”) may fall under Identity directed abuse.
- Some terms are sometimes pejorative, sometimes not, depending on the context so be cautious when identifying them. For example, ‘she is such a bitch!’ uses ‘bitch’ as a pejorative ‘my bitches and me!’ probably doesn’t. We distinguish between these different uses:
    - **Derogatory**: when the term is used to insult or attack someone – e.g. ‘That bitch sucks!!’

    - **Reclaimed**: When slurs are used by people from the targeted groups and become markers of solidarity and camaraderie rather than abuse
    - OR by an ally or someone clearly expressing support. This does not include so-called 'playful' use of slurs. – e.g. 'Beyoncé is such a bad bitch! Love her!!!!"
    - **Neutral**: when used in a general sense, without gendered or abusive connotations. This will only be possible for some pejorative terms – e.g. 'My bitches and me!' or 'Happy birthday you wee cunt!'
    - **Quoted**: when the pejorative is only used in a quote. This will be quoting another Reddit post, or from another context.

- Look out for new terms which include a more familiar misogynistic pejorative (e.g. "instasluts" contains "sluts"; "onlythots" contains "thots"). If these terms are used pejoratively against a person or group of people they should be annotated as misogynistic slurs. We are especially interested in discovering new terms in this category.
- When annotating pejoratives be very specific with your highlighting. You only need to highlight the pejorative word or term (e.g. "hoe" in "She's a hoe"). If the content has multiple pejoratives you should annotate them separately with distinct highlighting (e.g "beckies and stacies" should have separate annotations for "beckies" and "stacies"). **This is different from the other categories where we do want to keep the context of the abusive.**
- We are not including in this category pejoratives used to emasculate men by referring to their relationship with women (e.g. "cuck").
- If the content would not be abusive without the use of the pejorative, you can move onto the next piece of content. If the content would still be abusive without the pejorative, continue to label it using the other categories.
    - For example, 'My ex-girlfriend was a slut' should only be labelled for using the pejorative term 'slut'. However, 'my ex-girlfriend fucks any guys she sees – she's such a slut!' is also person-directed abuse. 'Women are such sluts. They need to be paid a lesson' should also be annotated as identity-directed.

2. Person-directed abuse

- Content which directs abuse against an *identifiable* person, who is either part of the conversation thread or is explicitly named in the conversation.
- The object of the abuse may be identified by their name – e.g. 'Helen Bonham Carter is a useless fucking cunt' – or their relationship to the author – e.g. 'my girlfriend is a cunt' – or, if they are part of the conversation thread, by their username – e.g. [responding to *random_username*] 'you are a cunt'
- In most cases, we don't know the personal relationships of people who we interact with online; assume that the content you annotate is shared by people who do not know each other. If you identify an entry that appears to be abusive but you think may actually be a joke between friends then annotate it as abusive but flag it in the comments section.
- Person-directed abuse involves two important annotations. First, is the relational status of the person being abused. This is divided into two sub-types:
    1. Abuse about a person. Content which directs abuse at a person who is not a participant in the conversation thread. The person must be identifiable to the

people in the conversation (i.e. there must be a genuine sense that the person under discussion is real and that the people talking about the person know them). This is primarily identified in the text by the person having their name written or by being tagged (e.g. by using @ or the /u/ flag). For example, ‘Helen Bonham Carter is a useless fucking cunt’.

a. Referring to someone by their *relationship* to you would qualify as abuse about a person even if they are not also named. For example, statements such as ‘My mum is a bitch’ or ‘I bloody hate my ex-girlfriend, she screwed me over’ would qualify as abuse about a person.

b. It is very rare that someone will be named who is not a prominent person or another Reddit user (again – you will know this because users are identified through clear symbols, such as /u/[user_name] then it counts as abuse about a person). If that happens – i.e. someone is named but you don’t know they are (even after a quick Google search) – then you should still annotate it up as Abuse about a person.

c. If you are unsure, flag the entry in the comments for discussion with the expert facilitator.

2. Abuse to a person. Content which directs abuse at a person who is part of the conversation thread, usually by directly attacking/insulting them or addressing a highly aggressive statement at them, such as ‘I hate you, you stupid bitch’, ‘u/User has no clue what she’s talking about, the daft bint and ‘Fuck off’.

a. This differs from abuse ‘about a person’ on the basis that the victim of the abuse is part of the conversation thread, i.e. they are someone who has already made an entry.

- For both abuse about and abuse to a person, make a note of the identity of the person (or the relationship to the poster if not known) in the comments.
- In addition, we want to know the gender of the target of abuse. This can be: woman, man, other, or unknown. If other, for example if the target is gender fluid, make a note in the comments.
- Finally, we want to know if the content of the abuse is misogynistic. For example, ‘Hilary Clinton has no clue what she’s talking about, the daft twat’ is abuse about a woman, but the content isn’t misogynistic. ‘Hilary Clinton is such a stupid bitch, someone should give her a good fucking and put her place’, however, does contain misogynistic abuse.

3. Identity-directed abuse

- Content which directs abuse at an identity or group. Note that the designation of a ‘group’ is open-ended and is not based on legal constraints (e.g. protected characteristics in the UK) or what platforms moderate for. *However, for this research we are only interested in abuse that is misogynistic*
- Multiple facets of identity can be targeted in one entry, such as ‘fucking women cause all our problems, and big darkies like Diane Abbot are the worse.’ In such cases, make a note in the comments section about the subgroup that is targeted (e.g. ‘black women’). We will then, afterwards, be able to investigate intersectional expressions of abuse.
- We distinguish between two categories of identity-directed abuse against women: treatment and derogation.

- Some forms of identity-directed abuse can be coded as implicit or explicit abuse.

3.a. Treatment

- Discussing negative treatment of women, either in the past, present, or future. This includes actions taken against women, as well as desires about how they should be treated. Treatment is broken down into two subcategories:
    - **Threatening language**: language which expresses an intent/desire to inflict/cause a targeted group to suffer harm, or expresses support for, encourages, advocates or incites such harm. These cases are likely to be explicitly abusive
        - The conceptual basis of 'threatening language' is ACTION. If you see a verb and/or an expression of intent (e.g. 'I will…', 'I want to…', 'I am going to….' or a normative statement that *others* should (e.g. 'They should be…' or 'We should') then you are likely looking at threatening language.
        - BUT remember at all times – there must be a group that is 'under threat' for an entry to be considered threatening language. If a user generically uses threatening language (such as 'I want to shoot everyone') then it is certainly concerning but, if a group is not identified, does not fall within this category.
        - If one entry expresses a threat ('I want to shoot some Muslims') and later entries express support for it (e.g. 'Yes mate, go do it!' or 'Right on!! You know it.') then they should also be marked up as threats. For the entries which are expressing support, it should be very unambiguous, as with these examples.
        - We have identified three themes of threatening language:
            - **Physical violence**: advocates non-sexual physical violence such as killing, maiming, beating, etc. e.g. 'Feminists deserve to be shot'
            - **Sexual violence**: advocates explicitly sexual violence such as rape, penetration, molestation, etc. e.g. 'Someone should rape her – that would put her in her place'
            - **Privacy**: advocates the invasion of privacy. e.g. 'I know where you live, bitch'
    - **Disrespectful actions**: Treating women as not independent, autonomous individuals. This includes more implicitly abusive statements about how women should be treated, or what they should be allowed to do.
        - Annotations in this category should also be flagged for whether they are explicitly or implicitly abusive
        - We have identified the following themes for controlling treatment of women:
            - **Manipulation**: this would include cases such as gaslighting, or lying to women to get them to do what you want
                - Implicit: 'Never tell women I'm a nurse if I want their respect'

        - Explicit: ‘Told my last girlfriend she was hallucinating when she saw the texts from my side piece’
    - **Seduction** & **Conquest**: discussing woman as sexual conquests, or describing previous incidences of when you have treated them as such.
        - Implicit: ‘Anyone got good tips for getting her from the club to the bed?’
        - Explicit: ‘Got her home and used her so hard’
    - **Controlling**: suggesting or stating that women should be controlled in some way.
        - Implicit: ‘I was not happy with the way she behaved’
        - Explicit: ‘I would never let my girlfriend do that’
- If you believe an entry includes abusive treatment of women that is not covered by the themes already described make a not in the comments and we will discuss it during facilitation.

3.b Derogation

- Content which explicitly derogates or demeans women. Most of this content will be descriptive, i.e. it describes how the author perceives things to be or express an opinion about how things are. We have defined two categories for derogation:
    - **Behaviour**: derogating women based on the way they behave *or should* behave. We have identified key themes:
        - **Financial ability**: this could include cases about women being bad with money, are unable to earn high wages
            - Implicit: ‘Child benefits are basically single mother welfare the government refuses to pay for’
            - Explicit: ‘We have a gender pay gap because woman don’t know how to negotiate’
        - **Careers**: this could demean women for having stereotypically feminine jobs, such as a kindergarten teacher, or belittle them for striving for stereotypically male jobs, such as those in STEM.
            - Implicit: ‘These women who judge me for being a nurse are usually social media managers or hostesses or some shit’
            - Explicit: ‘I work in tech and I’m sorry to say it but women’s brains just don’t work in the right way for the business’
        - **Traditional gender roles**: could suggest that women should fulfill certain roles, such as caretaking, homemaking, etc.
            - Implicit: ‘I couldn’t stay home with the kids – my wife would have to do it’
            - Explicit: ‘Women should get back in the kitchen and stay there’
    - **Attributes**: derogating women based on inherent parts of their character or body. We have identified three key themes:
        - **Intellectual inferiority:** judgements about women’s intellectual abilities such as a lack of critical thinking or emotional control. This will also include infantilizing women

        - Implicit: ‘My gf cries at the stupidest shit – lol!’
        - Explicit: ‘Typical stupid bitch – talking about things she doesn’t understand’
    - **Moral inferiority:** suggesting women are morally deficient in some way. Possible sub-themes include superficiality (e.g. only liking men who are rich or attractive), promiscuity, or untrustworthiness.
        - Implicit: ‘Girls love your money more than you’
        - Explicit: ‘My ex-girlfriend was a whore, she slept with every guy she saw’
    - **Sexual or physical limitations**: This can range from unattractiveness (i.e. lack of sexual desirability, ugliness (i.e. lack of beauty), frigidness (i.e. lack of sexual willingness) to subjective statements about feminine physical weakness**.** This would not include objective truths – e.g. woman don’t have penises so can’t be erections. If would include subjective opinions stated as facts – e.g. ‘women aren’t good at football’
        - Implicit: ‘I gave it my A-game but she would not give in!’
        - Explicit: ‘Yikes, Dianne Abbott looks like a monkey!’ (also intersectional)
- If abuse is derogatory but does not fit into the themes described use the flag **other** and make a note of the kind of abuse in the comments – ‘I don’t like women’ would be an example of implicit derogation while ‘All woman should shut the fuck up’ would be an explicit example. Both could be labelled as derogation ‘other’.

4 Counter-speech

- Counter-speech is content which challenges, condemns or calls out the abusive language of others. Counter-speech must be a RESPONSE to an existing piece of content in a thread. For instance, if an opening post attacks/calls out someone in another (e.g. offline) setting for what they have said then it is not counter speech (even if it is ‘countering’ what that person originally said). The content is only considered counter speech if it responds to what has been previously posted.
- Counter speech can take several forms, which include:
    1. Directly attack/condemn the abusive language in unambiguous terms, e.g. ‘I can’t believe you have just said that’ Or ‘You should not say things like that’ Or ‘that is just total nonsense.’
    2. Call out the original entry as abusive/hateful, e.g. ‘that is seriously prejudiced’
    3. Offer an alternative viewpoint which is clearly meant to challenge and undermine the original post (rather than just engage in ‘debate’ about abuse), e.g. ‘That is not at all what I have experienced and I think you’re completely wrong about this.’
    4. Attacking the author for what they have said, e.g. ‘You are a dickhead for sharing that sort of nonsense.’ Note that his would also qualify as person-directed abuse, ‘against a person’, and should also be flagged as such.

        - Mocking the original author or being sarcastic largely does NOT count as counter speech; if it is expressed in a light-hearted manner and does not seriously attack/criticise the author or their viewpoint. Such as: ‘Enlightening point from /r/user there!’
    5. Challenging the conclusions of the original author in a clear and unambiguous way, such as ‘That isn’t right. You don’t have the evidence to say that’.
- For an entry to be considered counter-speech the author must NOT engage in another form of abuse *against the same identity*. For instance, if the first post states, ‘Women are all sluts who want to control men’ and the second post states, ‘Woah, you shouldn’t be saying that!’ then the second post would be counter speech. However, if the second post then continues, ‘Women are not all sluts, women are just control freaks who want to ruin men’s lives!’ it would NOT be counter speech as the second author is only “challenging” the abuse of the first author to then engage in an only-slightly-different form of abuse themselves. In this case, the second author has not ‘rejected the expression of abuse’ but, rather, has refined it.
- We could end up with long chains of abuse and counter speech if two groups are being discussed and attacked/defended in alternating sequence. This is completely fine! Indeed, such chains are very interesting for our analysis.
- In many cases, counter speech may quote or reference abusive language. It should be clear that the author is only doing so in order to make an attack. Content which simply quotes a post and expresses shock/surprise/incredulity should not be viewed as counter-speech, e.g. using lots of punctuation (e.g. ‘?!?!?!!?’ after an abusive post) or using emojis (e.g. ‘:p’ after an abusive post). In such cases, it is still *ambiguous* as to whether the author is engaging in counter speech – and we want to be very sure that what we identify as counter-speech really is counter-speech.