An Expert Annotated Dataset for the Detection of Online Misogyny

Ella Guest
The Alan Turing Institute
University of Manchester
ella.guest@manchester.ac.uk

Bertie Vidgen
The Alan Turing Institute
bvidgen@turing.ac.uk

Alexandros Mittos
Queen Mary University of London
University College London
alexandros@mittos.net

Nishanth Sastry
University of Surrey
King's College London
The Alan Turing Institute
nsastry@turing.ac.uk

Gareth Tyson
Queen Mary University of London
The Alan Turing Institute
g.tyson@qmul.ac.uk

Helen Margetts
The Alan Turing Institute
Oxford Internet Institute
hmargetts@turing.ac.uk

Abstract

Online misogyny is a pernicious social problem that risks making online platforms toxic and unwelcoming to women. We present a new hierarchical taxonomy for online misogyny, as well as an expert labelled dataset to enable automatic classification of misogynistic content. The dataset consists of 6,567 labels for Reddit posts and comments. As previous research has found untrained crowdsourced annotators struggle with identifying misogyny, we hired and trained annotators and provided them with robust annotation guidelines. We report baseline classification performance on the binary classification task, achieving accuracy of 0.93 and F1 of 0.43. The codebook and datasets are made freely available for future researchers.

1 Introduction

Misogyny is a problem in many online spaces, making them less welcoming, safe, and accessible for women. Women have been shown to be twice as likely as men to experience gender-based online harassment (Duggan, 2017). This misogyny can inflict serious psychological harm on women and produce a 'silencing effect', whereby women self-censor or withdraw from online spaces entirely, thus limiting their freedom of expression (Mantilla, 2013; International, 2017). Tackling such content is increasingly a priority for social media platforms and civil society organisations.

However, detecting online misogyny remains a difficult task (Hewitt et al., 2016; Nozza et al., 2019). One problem is the lack of high-quality datasets to train machine learning models, which would enable the creation of efficient and scalable automated detection systems (Anzovino et al., 2018). Previous research has primarily used Twitter data and there is a pressing need for other platforms to be researched Lynn et al. (2019a). Notably, despite social scientific studies that show online misogyny is pervasive on some Reddit communities, to date a training dataset for misogyny has not been created with Reddit data. In this paper we seek to address the limitations of previous research by presenting a dataset of Reddit content with expert labels for misogyny that can be used to develop more accurate and nuanced classification models.

Our contributions are four-fold. First, we develop a detailed hierarchical taxonomy based on existing literature on online misogyny. Second, we create and share a detailed codebook used to train annotators to identify different types of misogyny. Third, we present a dataset of 6,383 entries from Reddit. Fourth, we create baseline classification models based on these datasets. All of the research artefacts are made freely available via a public repository for future researchers.[1]

The dataset itself has several innovations which differentiate it from previous training datasets for misogyny. First, we use chronological and structured conversation threads, which mean annotators take into account the previous context of each entry before labelling. Second, we distinguish between conceptually distinct types of misogynistic abuse, including gendered personal attacks, use of misogynistic pejoratives, and derogatory and threatening language. Third, we highlight the specific section of text, also known as a 'span', on which each label is based. This helps differentiate between multiple labels on one piece of text. Fourth, we use trained annotators, rather than crowd-sourced workers. We also use facilitated meetings to decide the final labels rather than just a majority decision. Both of these factors lead to a high-quality dataset. Additionally, we provide a second dataset with the original labels made by annotators before the final labels were decided.

[1] https://github.com/ellamguest/online-misogyny-eacl2021

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1336–1350
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

2 Background

Most previous classification work on online misogyny has used data from Twitter (Waseem and Hovy, 2016; Anzovino et al., 2018; Jha and Mamidi, 2017). However, social scientific and ethnographic research shows that Reddit is increasingly home to numerous misogynistic communities. Reddit is a social news website organised in to topic-based communities. Each subreddit acts as a message board where users make posts and hold discussions in comment threads on those posts. In recent years it has become a hub for anti-feminist activism online (Massanari, 2017; Ging and Siapera, 2018). It is also home to many misogynistic communities, particularly those associated with the 'manosphere', a loosely connected set of communities which perpetuate traditional forms of misogyny and develop new types of misogynistic discourse which in turn spread to other online spaces (Ging, 2017; Zuckerberg, 2018; Ging et al., 2019; Farrell et al., 2019; Ribeiro et al., 2020). Recent research suggests that the rate of misogynistic content in the Reddit manosphere is growing and such content is increasingly more violent (Farrell et al., 2019).

Waseem and Hovy (2016) provided a widely-used dataset for abusive language classification. They used expert annotators to identify sexist and racist tweets based on a set of criteria drawn from critical race theory. The tweets were initially labelled by the authors then reviewed by a third annotator. The resulting dataset consists of 17k tweets, of which 20% are labelled as sexist. However 85% of the disagreements between annotators were over sexism labels, which shows that even experienced coders of abusive language can have difficultly identifying gendered abuse.

Jha and Mamidi (2017) extended on the Waseem and Hovy (2016) dataset to distinguish between between 'benevolent' and 'hostile' sexism (Glick and Fiske, 1997). They classed all sexist labels in the previous dataset as 'Hostile' and all non-sexist labels as 'Other'. They then augmented the dataset by collecting tweets using keyword sampling on benevolently sexist phrases (e.g. 'smart for a girl') and extracted those manually identified as 'benevolent sexism'. In the combined dataset of 10,095 unique tweets 712 were labelled as 'benevolent', 2,254 as 'hostile', and 7,129 as 'not sexist'. They thus found that in the data hostile sexism was more than three times as common as the benevolent form. Their work highlights the need for greater attention to be placed on forms of 'subtle abuse', particularly for online misogyny (Jurgens et al., 2019).

Anzovino et al. (2018) developed a taxonomy with five categories of misogyny, drawn from the work of Poland (2016): Stereotype & Objectification, Dominance, Derailing, Sexual Harassment & Threats of Violence, Discredit. They used a combination of expert and crowdsourced annotation to apply the taxonomy and present a dataset of 4,454 tweets with balanced levels of misogynistic and non-misogynistic content. A shared task confirmed that the dataset could be used to distinguish misogynistic and non-misogynistic content with high accuracy, but performance was lower in differentiating between types of misogyny (Fersini et al., 2018).

Lynn et al. (2019b) provide a dataset of 2k Urban Dictionary definitions of which half are labelled as misogynistic. In Lynn et al. (2019a) they show that deep learning techniques had greater accuracy in detecting misogyny than conventional machine learning techniques.

3 Data collection

We collected conversation threads from Reddit. Given that a very small amount of content on social media is hateful, a key difficulty when creating datasets for annotation is collecting enough instances of the 'positive' class to be useful for machine learning (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). However, sampling strategies can introduce biases in the composition and focus of the datasets if overly simplistic methods are used, such as searching for explicitly misogynistic terms (Wiegand et al., 2019).

To ensure that our dataset contains enough misogynistic abuse we began with targeted sampling, taking content from 12 subreddits that were identified as misogynistic in previous research. This includes subreddits such as `r/MensRights`, `r/seduction`, and `r/TheRedPill`. The sources used to identify these subreddits are available in Table 9 in the Appendix. We then identified 22 additional subreddits which had been recommended by the moderators/owners of the original 12 subreddits in the 'sidebar'. Some of these are not misogynistic but discuss women (e.g. `r/AskFeminists`) and/or are otherwise related to misogyny. For example, `r/exredpill` is a support group for former members of the misogynistic subreddit `r/TheRedPill`. Table 9 in the Appendix lists the 34 targeted subreddits and the number of entries and threads for each in the dataset. Over 11 weeks, for each subreddit, we collected the entire threads of the 20 most popular posts that week.

Using subreddits to target the sampling rather than keywords should ensure that more linguistic variety is captured, minimising the amount of bias as keywords such as 'slut' are associated with more explicit and less subtle forms of abuse. Nonetheless, only sampling from suspected misogynistic communities could still lead to classifiers which only identify the forms of misogyny found in those targeted contexts (Davidson et al., 2017; Wiegand et al., 2019; Sap et al., 2019). To account for this potential bias, and to enable greater generalisabilty, we sampled content from 71 randomly selected subreddits. They accounted for 18% of threads and 16% of entries in our dataset. For each randomly selected subreddit, we collected the thread of the most popular post. All threads were in English with the exception of one thread from the subreddit `r/Romania`.

Posts and comments were collected from February to May 2020 using the python package PRAW, a wrapper for the Reddit API (Boe, 2020). Posts on Reddit have a text title and a body which can be text, an image, or a link. For posts with a text body we combined this with the post title to create a single unit of text. For the 29% of posts where the body was an image we also collected the image.

4 Taxonomy

We developed a hierarchical taxonomy with three levels. First, we make a binary distinction between Misogynistic content and Non-misogynistic content, which are mutually exclusive. Second, we elaborated subtypes of Misogynistic and Non-misogynistic content. For Misogynistic content we defined four categories: (i) *Misogynistic Pejoratives*, (ii) descriptions of *Misogynistic Treatment*, (iii) acts of *Misogynistic Derogation* and (iv) *Gendered Personal attacks* against women. For Non-misogynistic content we defined three categories: (i) *Counter speech* against misogyny, (ii) *Non-misogynistic personal attacks* and (iii) *None of the categories*. Third, we included additional flags for some of the second level categories. Within both Misogynistic and Non-misogynistic content, the second level categories are not mutually exclusive, thereby allowing for multiple labels per entry. For instance, a Misogynistic entry could be assigned labels for both a *Pejorative* and *Treatment*.

This taxonomy draws on the typologies of abuse presented by Waseem et al. (2017) and Vidgen et al. (2019) as well as theoretical work in online misogyny research (Filipovic, 2007; Mantilla, 2013; Jane, 2016; Ging, 2017; Anzovino et al., 2018; Ging and Siapera, 2019; Farrell et al., 2019). It was developed by reviewing existing literature on online misogyny and then iterating over small samples of the dataset. This deductive-inductive process allowed us to ensure that conceptually distinct varieties of abuse are separated and that different types of misogyny can be unpicked. This is important given that they can have very different impacts on victims, different causes, and reflect different outlooks and interests on the part of the speaker.

4.1 Misogynistic content

Misogynistic content directs abuse at women or a closely related gendered group (e.g. feminists). This content can fall in to four non-mutually exclusive categories.

4.1.1 Misogynistic pejoratives

Misogynistic pejoratives are terms which are used to disparage women. It includes terms which are explicitly insulting and derogatory, such as 'slut' or 'whore', as well as terms which implicitly express negativity or animosity against women, such as 'Stacy' or 'Becky'. For example, 'Stacy' is a term used in the incel community to describe women considered attractive and unattainable, in opposition to a more average and attainable 'Becky' (Jennings, 2018).

4.1.2 Misogynistic treatment

Misogynistic treatment is content that discusses, advocates, incites or plans negative or harmful treatment of women. It includes expressing intent to take action against women, as well as expressing desires about how they should be treated. Misogynistic treatment contains third-level subcategories: *Threatening language* and *Disrespectful actions*.

1. *Threatening language*: Content which expresses an intent/desire to inflict/cause women to suffer harm, or expresses support for, encourages, advocates or incites such harm. It is an 'explicit' form of abuse. It falls in to three thematic groups:

(a) *Physical violence*: non-sexual physical violence such as killing, maiming, beating, etc. e.g. 'Feminists deserve to be shot'.

(b) *Sexual violence*: explicit sexual violence such as rape, penetration, molestation, etc. e.g. 'Someone should rape her – that would put her in her place'.

(c) *Privacy*: an invasion of privacy such as the disclosure of personal information (i.e. doxing) or threats to visit them. e.g. 'I know where you live, bitch'.

2. *Disrespectful actions*: Content which treats/portrays women as either lacking or not deserving independence/autonomy. This includes more subtly abusive statements about how women should be treated and what they should be allowed to do. It is an 'implicit' form of abuse. It falls in to four thematic groups:

   (a) *Controlling*: suggesting or stating that women should be controlled in some way, especially by a man or men. E.g. 'I would never let my girlfriend do that'.

   (b) *Manipulation*: using or advocating the use of tactics such as lying and gaslighting to manipulate what women do or think. E.g. 'Told my last girlfriend she was hallucinating when she saw the texts from my side piece'.

   (c) *Seduction and conquest*: discussing woman solely as sexual conquests or describing previous incidences of when they have been treated as such. E.g. 'Got her home and used her so hard'.

   (d) *Other*: content that is not covered by the other subcategories.

4.1.3 Misogynistic derogation

Misogynistic derogation is content that demeans or belittles women. This content can be explicitly or implicitly abusive. It is separated into third-level subcategories:

1. *Intellectual inferiority*: making negative judgements of women's intellectual abilities, such as a lack of critical thinking or emotional control. This includes content which infantilizes women. An implicit example would be 'My gf cries at the stupidest shit – lol!' for suggesting irrational emotional responses. An explicit example is 'Typical stupid bitch – talking about things she doesn't understand'.

2. *Moral inferiority*: making negative judgements of women's moral worth, such as suggesting they are deficient or lesser to men in some way. This includes subjects such as superficiality (e.g. only liking men who are rich or attractive), promiscuity, and untrustworthiness. An implicit example is 'Girls love your money more than you'. An explicit example is 'My ex-girlfriend was a whore, she slept with every guy she saw'.

3. *Sexual and/or physical limitations*: making negative judgements of women's physical and/or sexual ability. This includes perceived unattractiveness (i.e. a lack of sexual desirability), ugliness (i.e. a lack of beauty), frigidness (i.e. a lack of sexual willingness), as well as belittling statements about feminine physical weakness. An implicit example is 'I gave it my A-game but she would not give in, so uptight!' An explicit example is 'Yikes, Dianne Abbott looks like a monkey!'

4. *Other*: content that is not covered by the other subcategories but is derogatory towards women.

4.1.4 Gendered personal attacks

Gender personal attacks are highly gendered attacks and insults. This category is used only when the nature of the abuse is misogynistic, e.g. 'Hilary Clinton is such a stupid bitch, someone should give her a good fucking and put her in her place'.

The category has a level three flag for the *gender of the recipient* of the abuse. We include this flag as research has shown that men can also be targeted by misogynistic attacks (Jane, 2014). The gender can either be a woman (e.g. 'That chick is dumb'), a man (e.g. 'This dude is a piece of shit') or unknown (e.g. 'You're are an idiot, fuck off'). If the content was replying to an entry which reveals the recipient's gender we can infer it from this context. For example if 'You're an idiot, fuck off' was a response to 'I'm a man and a feminist there's nothing contradictory about that' we know the abuse is targeted at a man.

4.2 Non-misogynistic content

Non-misogynistic content can fall in to three non-mutually exclusive categories, all of which are relevant for misogyny research.

4.2.1 Non-misogynistic personal attacks

Interpersonal abuse which is *not* misogynistic. We include this category to allow for a comparison of the nature of abuse directed at women and men (Duggan, 2017). It includes content which personally attacks a woman but is not misogynistic in nature, e.g. 'Hilary Clinton has no clue what she's talking about, idiot!'. It uses the same level three flag for the *gender of the recipient* as Misogynistic personal attack. This allows us to compare the rates of personal attacks against women and men.

Note that although it is possible for an entry to contain both Misogyny and a Non-misogynistic personal attack, this was very rare. In such cases, we chose to not annotate the Non-misogynistic personal attack in order to keep the first level as a binary distinction.

4.2.2 Counter speech

Counter speech is content which challenges, refutes, and puts into question previous misogynistic abuse in a thread. It could directly criticise previous abuse (e.g. 'What you said is unacceptable'), specifically accuse it of prejudice (e.g. 'That's incredibly sexist'), or offer a different perspective which challenges the misogyny (e.g 'That's not how women act, you're so wrong').

4.2.3 None of the categories

Content which does not contain misogynistic abuse, pejoratives, or related counter speech as defined in the previous categories. This content is often not related to abuse or to women in general. That said, it can include other forms of abusive language which are not misogynistic.

5 Annotation Methodology

A key difficulty in the formation of abusive language training datasets is producing high quality annotations. Several factors affect this. Deciding between similar categories, such as 'hate speech' versus 'offensive language' can be difficult (Waseem et al., 2017). Determining the right category often requires close scrutiny and sustained critical thinking from annotators. Annotators may face information overload if asked to work with too many categories, both in terms of breadth (e.g. annotating for different types of abuse) and depth (e.g. working with numerous subcategories). Further, annotators may have different values and experiences and so make different assessments of the content they observe, especially when context plays a large role. Annotators will also have unconscious social biases which may mean they interpret coding instructions differently to each other, and to how they were intended by the research authors. For instance, Davidson et al. (2017) found that crowdsourced annotators were more likely to label sexist content as merely 'offensive' while racist and homophobic content was considered 'hate speech'.

To mitigate such annotator biases, we used expert annotators specifically trained in identifying misogynistic content, as well as a group-based facilitation process to decide final labels. Due to time and resource constraints, the final dataset is smaller than if we had used crowdsourced workers but captures more nuanced and detailed cases of misogyny. Six annotators worked on the dataset. Annotators were trained in the use of a codebook detailing the taxonomy and annotation guidelines. The codebook was updated over time based on feedback from the annotators. Demographic information on the annotators is available in Appendix A.2

5.1 Annotation process and disagreements

Annotators independently marked up each entry for the three levels presented in Section 4. For all level two categories other than 'None', they also highlighted the specific part of the entry which was relevant to the labelled category (the 'span'). This is particularly important information for long posts which can contain multiple forms of abuse.

Each entry was annotated by either two (43%) or three (57%) annotators. If all annotators made the exact same annotation (including all three levels and highlighting) this was accepted as the final annotation. All other entries were flagged as disagreements. Annotators reviewed the disagreements in weekly meetings which were overseen by an expert facilitator, a PhD researcher who had developed the annotation taxonomy and was familiar with the literature on online misogyny and hate speech classification. The role of the facilitator was to promote discussion between annotators and ensure the final labels reflected the taxonomy. Each disagreement was discussed until the annotators reached a consensus on the final agreed label or labels.

5.2 Inter-annotator reliability

For the level one binary task the Fleiss' Kappa is 0.484 and the Krippendorf's alpha is 0.487. By conventional NLP standards these results appear low. However they are equivalent to, or above, those of existing abusive content datasets. Sanguinetti et al. (2018) report category-wise Kappas from k=0.37 for *offence* to k=0.54 for *hate*. Gomez et al. (2020) have a Kappa of 0.15 in the "MMH150" dataset of hateful memes. Fortuna and Nunes (2018) report a Kappa of 0.17 for a text-only task. Krippendorf's alpha is similar to the 0.45 reported by Wulczyn et al. (2017).

We also calculated level two category-wise Fleiss' Kappas for each of the 17 sets of annotator groups, then took the mean across all groups (Ravenscroft et al., 2016). Table 1 shows the breakdown of Kappas per category. There was greatest agreement for *Misogynistic pejoratives* (k=0.559) down to the lowest agreement for *Misogynistic personal attacks* (k=0.145).

Category	Fleiss' Kappa
Mis. Pejoratives	0.559
Mis. Treatment	0.210
Mis. Derogation	0.364
Mis. Personal attack	0.145
Nonmis. Personal attack	0.239
Counter speech	0.179
None of the categories	0.485

Table 1: Category-wise Fleiss' Kappa

Our taxonomy has seven partially overlapping categories, and as such annotation is considerably more difficult compared with most prior work, which tends to involve only binary labelling. As such, whilst slightly low, we believe that our agreement scores show the robustness of our annotation approach. Further, all disagreements were then discussed with an expert adjudicator, meaning that points of disagreement were addressed before the final labels were determined.

6 Prevalence of the categories

Of the 6,567 agreed labels in the final dataset 10.6% are Misogynistic (n=699) and 89.4% are Non-misogynistic (n=5,868). Tables 2 and 3 show the number of labels in the final dataset for each of the Misogynistic and Non-misogynistic categories, broken down by the level two categories. The vast majority of entries fall under *None of the categories* (88.6% of all labels). The next most common category is *Misogynistic Pejoratives* followed by *Misogynistic Derogation* pejoratives (4.2%). There are relatively few labels for *Personal attacks* with just 0.7% in total for each of the Misogynistic and Non-misogynistic categories, respectively. The least common category is *Counter speech* against misogyny, with only ten cases (0.2%).

Category	Number	Total %
Pejorative	276	4.2%
Treatment	103	1.6%
Derogation	285	4.3%
Personal attack	35	0.7%
Total	**696**	**10.6%**

Table 2: Breakdown of *Misogynistic* category counts

Category	Number	Total %
Personal attack	43	0.7%
Counter speech	10	0.2%
None	5815	88.6%
Total	**5868**	**89.4%**

Table 3: Breakdown of *Non-misogynistic* category counts

6.1 Misogynistic pejoratives

Annotators identified at least one misogynistic pejorative in 4.2% of all entries. The most common misogynistic term in the labels is 'bitch' (n=43) followed by 'stacy' (24) and 'stacies' (21).

6.2 Misogynistic treatment

There are 103 labels of *Treatment*. Figure 1 shows the number of labels for each level three subcategory. There are almost five times as many labels for *Disrespectful actions* (n=85) than *Threatening language* (n=18).

Both level three subcategories were broken down into more specific misogynistic themes. Within *Disrespectful actions*, *Seduction and conquest* is the most common topic, with twice as many labels as the second most common, *Controlling* (43 vs 17). And, within *Threatening language*, *Physical violence* was the most common theme (13) while

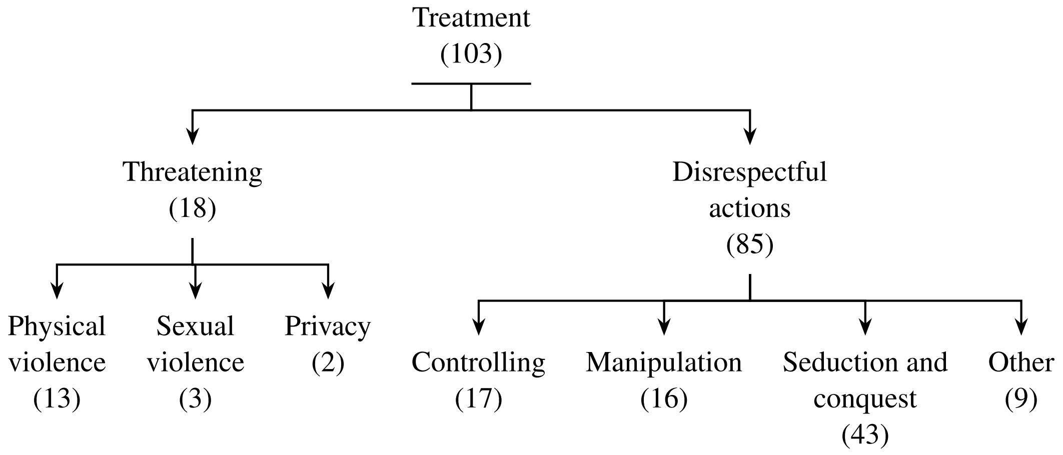

Figure 1: Prevalence of Misogynistic Treatment subcategories

Sexual violence and *Invasion of privacy* only have a couple of labels each (three and two, respectively).

6.3 Misogynistic derogation

The are 286 *Derogation* labels. Table 4 shows the number of labels for each subcategory within *Derogation*. The counts are broken down by the strength of the abuse (i.e. *implicit/explicit*). *Implicit derogation* is almost twice as common as *explicit* (182 vs 103). The most common subcategory is *Moral inferiority* which accounts for 51% of *implicit* and 54% of *explicit Derogation*. *Intellectual inferiority* then has equal numbers of *implicit* and *explicit* labels (n=16).

Subcategory	Implicit	Explicit
Moral infer.	92	56
Intellectual infer.	16	16
Sexual & physical lim.	14	12
Other	60	19
Total	**182**	**103**

Table 4: Breakdown of *Misogynistic Derogation* subcategories by implicit and explicit strength

6.4 Personal attacks

Table 5 shows the breakdown of both *Misogynistic* and *Nonmisogynistic* personal attacks. Slightly more than half (55%) of interpersonal abuse was *not* misogynistic. Of these women were still the target of the abuse almost four times as often as men (n=32 vs n=8). And women were as likely to receive misogynistic person attacks as non-misogynistic ones (n=32).

Gender	Misog.	Nonmis.	Total
Woman	32	32	64 (82%)
Man	2	8	10 (13%)
Unknown	1	3	4 (5%)
Total	35 (45%)	43 (55%)	78

Table 5: Breakdown of *Misogynistic* and *Nonmisogynistic* personal attacks by *Gender* of the target

The gender of the target was only *unknown* in 5% of cases, one misogynistic and three not. There were two cases of misogynistic abuse against men. All other misogynistic personal attacks were towards women.

6.5 Counter speech

There are only 10 cases of *Counter speech* in the final dataset of agreed labels. Annotators originally identified far more counter speech (188 labels for 149 unique entries were initially made) but few were accepted during the adjudication meetings. In Section 5.2 we showed that the category has one of the lowest Kappa values. Notably, 39% of original *Counter speech* labels were made by one annotator, showing that the annotators had different understandings of the threshold for *Counter speech*. However, the number of original labels for *Counter speech* decreased over the first few weeks of the annotation process, as shown in Figure 2. This reflects the complexity of the category; it took annotators time to differentiate content that was pro-women from that which actually countered previous misogynistic speech.

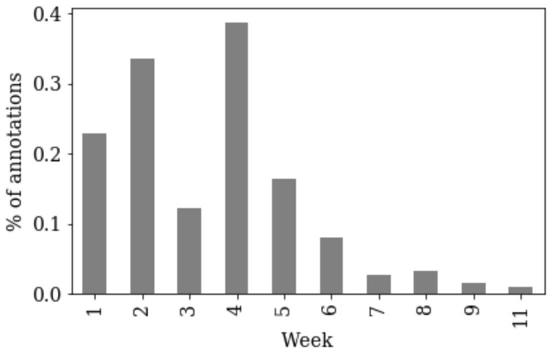

Figure 2: Percentage of original *counter speech* labels by week

7 Experiments

As reference points for further research using our dataset, we provide three experimental baselines on the binary task of distinguishing between misogynistic and non-misogynistic content, i.e. level one of our taxonomy. As the simplest baseline, we evaluate a logistic unigram classifier. Further, we evaluate two uncased BERT-base models (Devlin et al., 2019) – one unweighted, the other using class weights emphasising the minority class, i.e. misogynistic content, to account for class imbalance. For all models, we use the same stratified 80/20 train/test split of the dataset. Details on model training and parameters can be found in Appendix C.

Performance of the three models is shown in Table 6. All models perform poorly on misogynistic content, with the logistic classifier performing worst overall. The logistic classifier has the highest precision on misogynistic content (0.88) but very low recall (0.07) and a low F1 score (0.13). The weighted BERT model has the highest recall (0.50) and F1 score (0.43). Accuracy on all test cases, of which 91.9% are non-misogynistic, is around 0.90 across models.

The classification task is complicated by the relatively small size of our dataset (n=6,385 unique cases) as well as the relatively small proportion of misogynistic cases in it (8.1%). These issues are common in abusive speech detection (Fortuna and Nunes, 2018; Fortuna et al., 2020). To address them, future research can leverage the typology and annotation process we introduced to collect additional cases, particularly misogynistic ones, thus growing and balancing the dataset.

7.1 Error analysis

We make use of the more granular secondary labels in our taxonomy to conduct an error analysis for the weighted BERT model. Table 7 shows the confusion matrix for the 1,277 entries in the test set. Overall, 137 entries (10.7%) were misclassified. 86 (63%) of these errors were false positives and 51 (37%) were false negatives.

7.1.1 False positives

Of the 86 entries misclassified as *Misogynistic*, three are labelled as *Nonmisogynistic personal attacks*. An example is “Male feminist reacts to vile scumbag who murdered his kids by telling ALL men to, you know, stop killing women and kids...”. The use of aggressive language combined with gendered phrases (such as “stop killing women”) likely led to its misclassification.

The remaining 83 false positives fall under *None of the categories* and all contain some reference to women. Some refer to topics often associated with misogyny but are not misogynistic in themselves. For example, a comment in `r/seduction` stated, “the most manly thing is to find your dream woman, marry her, and live happily ever after. The constant sex with women is so overrated anyways”. This entry, which suggests that other things than high levels of sexual activity should be prioritised, is thematically similar to misogynistic content in the dataset.

Other false positives mention women indirectly. “Because they aren’t men, they are SIMPS”. ‘Simp’ is a pejorative term used in the manosphere for a man who cares too much about a woman. Under our taxonomy it did not count as a misogynistic pejorative but it is likely that the term appears in misogynistic entries in the dataset. Some false positives are critical of misogyny, though not actively enough to count as *Counter speech*. For example “Does this moid even know the meaning of the term ‘butterface’? If this woman is ugly, there is no hope for most of the female population.”. This discussion of unrealistic beauty standards of women references misogyny but is not itself misogynistic.

7.1.2 False negatives

Of the 51 Misogynistic entries the model misses, almost half (n=24) contain *Derogation*. Implicit and explicit derogation are missed at roughly similar rates, as are each of the subcategories. Importantly this shows that the different forms of derogation are no more or less likely to be missed.

Model	Precision	Recall	F1 score	Accuracy
Logistic regression	**0.88**	0.07	0.13	0.92
BERT (unweighted)	0.67	0.30	0.42	**0.93**
BERT (weighted)	0.38	**0.50**	**0.43**	0.89

Table 6: Model performance on misogynistic test cases (n=103) and accuracy on all test cases (n=1,277).

		Prediction		
		Nonmis.	Mis.	TOTAL
Label	Nonmis.	1,088	86	1,174
	Mis.	51	52	103
	TOTAL	1,139	138	1,277

Table 7: Confusion matrix of the weighted BERT model

In many cases, the derogation depends on the context of the earlier conversation thread, thus the BERT-model, which does not explicitly take into account prior entries in the thread, cannot recognise the misogyny in isolation. "It's funny to see the hamster that starts to act up in their little widdle tiny brains after saying that too." is an explicit statement that women are intellectually inferior, but understanding that it refers to women depends on having seen previous entries in the conversation.

The next most common source of false negatives is *Pejoratives* (n=19). The classifier misses six counts each of 'whore' and 'stacy' and five of 'bitch'. There are seven missed labels for *Treatment*, five *Disrespectful actions* and two *Threatening language*. However, due to the low prevalence of the category in the training data we anticipate some errors. For example, "I am waiting for such incident to happen to me so that I can beat the shit out of her, and of course it will be all revenge" details a specific form of violence (i.e. 'beat the shit out of her') which the model cannot know to identify as misogyny without being trained on other uses of the term.

The final two errors are for *Personal attacks*. For example, "Yeah theres women that I as an Incel wouldnt even acknowledge and this is one of em [sic]". This is an implicit attack which requires understanding that considering a woman unworthy of the attention of an incel is a gendered insult.

As we can see from these examples the main classification errors are due to context limitations. For false negatives there is usually not enough information in the entry alone or in the training dataset to identify the misogyny. Conversely, for false positives the classifier appears to overly associate content *about* women with content that *abuses* women. These limitations can be addressed by future work drawing on the taxonomy and annotation process presented here to develop larger datasets which can cover a greater range of forms of discourse, including both non-misogynistic discussions of women and a wider variety of misogynistic speech.

8 Conclusion

In this paper we have presented a hierarchical granular taxonomy for misogyny and have described a dataset containing high quality, expert labels of misogynistic content from Reddit. We have also provided the detailed coding book we created and a dataset with all of the original labels. The final dataset is small compared to other annotated datasets used for classification. However it benefits from a detailed taxonomy based on the existing literature focused on just one form of online abuse - misogyny. The use of trained annotators and an adjudication process also ensures the quality of the labels.

The more granular subcategories in the taxonomy may be too small to classify separately, but they provide insights into the relative frequency of different forms of misogynistic content on Reddit and enable detailed error analysis. They are also useful for other researchers aiming to create larger datasets, who can build on the taxonomic work conducted here.

Acknowledgments

This research was funded by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1 at The Alan Turing Institute, particularly the "Tools, Practices and System" and "Criminal Justice" themes.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 57–64, Cham. Springer International Publishing.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bryce Boe. 2020. PRAW. Python Reddit API Wrapper Development.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, page 4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Maeve Duggan. 2017. Online Harassment 2017. Technical report, Pew Research Center.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In *WebSci '19 Proceedings of the 10th ACM Conference on Web Science*, pages 87–96, Boston.

E Fersini, P Rosso, and M Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) 215*, page 15.

Jill Filipovic. 2007. Blogging while female: How internet misogyny parallels real-world harassment. *Yale JL & Feminism*, 19:295.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6786–6794.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Debbie Ging. 2017. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, 1:20.

Debbie Ging, Theodore Lynn, and Pierangelo Rosati. 2019. Neologising misogyny: Urban Dictionary's folksonomies of sexual abuse:. *New Media & Society*.

Debbie Ging and Eugenia Siapera. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18(4):515–524.

Debbie Ging and Eugenia Siapera, editors. 2019. *Gender Hate Online: Understanding the New Anti-Feminism*. Springer International Publishing, Cham.

Peter Glick and Susan T. Fiske. 1997. Hostile and Benevolent Sexism. *Psychology of Women Quarterly*, 21(1):119–135.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Snowmass Village, CO, USA. IEEE.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The Problem of Identifying Misogynist Language on Twitter (and Other Online Social Spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 333–335, New York, NY, USA. ACM.

Amnesty International. 2017. Amnesty reveals alarming impact of online abuse against women. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/.

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

Emma Alice Jane. 2014. 'Back to the kitchen, cunt': Speaking the unspeakable about online misogyny. *Continuum*, 28(4):558–570.

Rebecca Jennings. 2018. Incels Categorize Women by Personal Style and Attractiveness. https://www.vox.com/2018/4/28/17290256/incel-chad-stacy-becky.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019a. A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–8.

Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019b. Data set for automatic detection of online misogynistic speech. *Data in Brief*, 26:104223.

Karla Mantilla. 2013. Gendertrolling: Misogyny Adapts to New Media. *Feminist Studies*, 39(2):563–570.

Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence on - WI '19*, pages 149–155, Thessaloniki, Greece. ACM Press.

Bailey Poland. 2016. *Haters: Harassment, Abuse, and Violence Online*. Potomac Books.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. *arXiv:1909.04251 [cs]*.

James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. 2016. Multi-label Annotation in Scientific Articles - The Multi-label Cancer Risk Assessment Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4115–4123, Portorož, Slovenia. European Language Resources Association (ELRA).

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2020. From Pick-Up Artists to Incels: A Data-Driven Sketch of the Manosphere. *arXiv:2001.07600 [cs]*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In

Proceedings of the 26th International Conference on World Wide Web, pages 1391–1399, Perth Australia. International World Wide Web Conferences Steering Committee.

Donna Zuckerberg. 2018. *Not All Dead White Men: Classics and Misogyny in the Digital Age*. Harvard University Press, Cambridge, Massachusetts.

A Short form data statement

Following the recommendation of Bender and Friedman (2018) we include the following short form data statement to summarise the main features of the datasets. Further details on the creation of the datasets are in in Sections 3 and 5 in the main paper.

A.1 Data

The two datasets include labels for 6,383 unique Reddit entries (i.e. posts or comments) across 672 conversation threads collected. One dataset is of the 15,816 original labels selected by annotators and the second is of the 6,567 agreed labels. Table 8 provides a description of each of the variables in the datasets. We also include the accompanying set of images associated with some original post entries.

All threads except one are in English. The majority of threads were sampled from a set of 34 subreddits selected for the expected prevalence of misogynistic content, or non-misogynistic discussions about women. Paid annotators received extensive training to apply the taxonomy presented in this paper to label entries. The majority of annotators were White-British, spoke English as a first language, and had or were pursuing a University degree. Two-thirds of annotators were women.

A.2 Annotators

All annotators were based in the United Kingdom and worked remotely. They were paid £14 per hour for all work including training. Five of the six annotators gave permission to share their basic demographic information. All were between 18 and 29 years old. Two had high school degrees, two had an undergraduate degree, and one had a postgraduate taught degree or equivalent. Four identified as women, one as a man. All were British nationals, native English speakers, and identified as ethnically white.

All annotators used social media at least once per day. Two had never been personally targeted by online abuse, two had been targeted 2-3 times (in separate instances more than a year ago), and one had been personally targeted more than 3 times within the previous month.

B Frequency of targeted subreddits

Table 9 lists the subreddits used for target sampling of data. The columns *Num entries* and *Num threads* state how many individual entries and threads from each subreddit are in the datasets. The column *Selection* shows whether the subreddit was identified from existing literature, which is cited, or using snowball sampling.

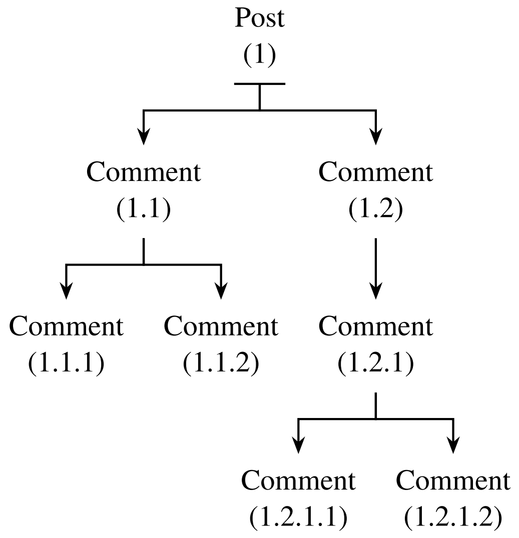

Figure 3: Tree diagram for comment order in threads

C Model Details

Pre-Processing We lowercase all text and remove newline and tab characters. URLs and emojis are replaced with [URL] and [EMOJI] tokens.

C.1 Logistic regression

Logistic regression with l1-regularisation is implemented in R using the 'glmnet' package (Friedman et al., 2010) on a unigram representation of the data. Lambda is selected using cross-validation and set to 0.015.

C.2 BERT Models (weighted/unweighted)

Model Architecture We implement uncased BERT-base models (Devlin et al., 2019) using the `transformers` Python library (Wolf et al., 2020). For sequence classification, we add a linear layer with softmax output.

Training Parameters We apply a stratified 80/20 train/test split to our dataset. Models are trained for three epochs each. Training batch size is 16. We use cross-entropy loss. For the weighted model, we add class weights emphasising the minority class, i.e. misogynistic content. Weights are set to the relative proportion of the other class in the training data, meaning that for a 1:9 misogynistic:non-misogynistic case split, loss on misogynistic cases would be multiplied by 9. The optimiser is AdamW (Loshchilov and Hutter, 2018) with a 5e-5 learning rate and a 0.01 weight decay. For regularisation, we set a 10% dropout probability.

Variable	**Description**
entry_id	A unique string assigned to every comment and post by Reddit.
link_id	The id number of the original post of a thread.
parent_id	The id number of parent entry (i.e. the post or comment this entry responds to).
subreddit	The subreddit community where the entry was made.
author	The Reddit username of the entry author.
body	The text body of the entry. For the original posts of threads the title and post body were combined.
image	Whether the entry has an accompanying image. Only applicable to posts. Images are provided as jpg files. They are named as 'X_Y_Z' corresponding to the week (X), group (Y), and thread id (Z).
label_date	The week commencing date of when the entry was labelled.
week	The week in the annotation process when the entry as assigned (1 to 11).
group	The weekly group the entry was assigned to. All weeks had two groups except week 7 which only had 1.
sheet_order	The order of the entry in the weekly annotation sheet. This is a list of numbers referring to the nested structure of comments in threads. It shows the id number of each level of the thread from the original post to the relevant entry. For example, if an entry has the sheet_order (1, 2, 3) it belongs to the first thread (1), and replied to the second comment (2), to which it is the third reply (3). See Fig. 3 for visual explanation.
annotator_id	The id number of the annotator who made the annotation (1 to 6). Only applicable to the original_labels dataset.
level_1	Whether the entry is *Misogynistic* or *Nonmisogynistic*.
level_2	The category of the label (i.e. *Pejoratives, Derogation*, etc.
level_3	EITHER the subcategory for *Derogation* or *Treatment* OR the gender of the target for either *Personal attack* category. Empty for all other categories.
strength	Whether the abuse is implicit or explicit. Only applicable to identity directed abuse.
highlight	The highlighted part of the entry's body which contains the abuse. Mandatory for all primary categories except 'None'.
split	Whether the entry was included in the 'train' or 'test' dataset split for model building. Only applicable to the final_labels dataset.

Table 8: Description of dataset variables

Subreddit	Num entries	Num threads	Selection
altTRP	2	1	Snowball
AskFeminists	263	26	Snowball
askseddit	142	16	Snowball
badwomensanatomy	430	31	Farrell et al. (2019)
becomeaman	2	1	Snowball
Egalitarianism	115	15	Snowball
exredpill	113	12	Snowball
FeMRADebates	195	20	Snowball
GEOTRP	11	1	Snowball
IncelsInAction	110	14	Farrell et al. (2019)
IncelsWithoutHate	325	28	Farrell et al. (2019)
KotakuInAction	373	28	Qian et al. (2019); Zuckerberg (2018)
marriedredpill	87	7	Snowball
masculism	34	5	Snowball
MensRants	4	1	Ging (2017)
MensRights	364	29	Ging (2017); Qian et al. (2019); Zuckerberg (2018)
mensrightslaw	2	1	Snowball
MensRightsMeta	4	1	Snowball
MGTOW	601	41	Farrell et al. (2019); Ging (2017); Qian et al. (2019); Zuckerberg (2018)
mgtowbooks	2	1	Snowball
MRActivism	8	2	Snowball
NOMAAM	2	1	Snowball
pua	10	1	Snowball
PurplePillDebate	221	21	Snowball
PussyPass	344	33	Qian et al. (2019)
pussypassdenied	262	22	Qian et al. (2019)
RedPillParenting	12	2	Snowball
RedPillWives	61	8	Snowball
RedPillWomen	217	23	Snowball
seduction	392	33	Zuckerberg (2018)
ThankTRP	8	1	Snowball
TheRedPill	338	29	Ging (2017); Zuckerberg (2018)
theredpillright	10	1	Snowball
Trufemcels	434	37	Farrell et al. (2019)

Table 9: Number of entries and threads per targeted subreddit