Identity Construction in a Misogynist Incels Forum

Michael Miller Yoder,[1] **Chloe Perry,**[2] **David West Brown,**[3]
Kathleen M. Carley,[1] **Meredith Pruden**[4]

[1]Software and Societal Systems Dept., Carnegie Mellon University, Pittsburgh, PA, USA
[2]Dept. of American Culture, University of Michigan, Ann Arbor, MI, USA
[3]Dept. of English, Carnegie Mellon University, Pittsburgh, PA, USA
[4]School of Communication and Media, Kennesaw State University, Kennesaw, GA, USA
yoder@cs.cmu.edu, chloeper@umich.edu, dwb2@andrew.cmu.edu,
carley@cs.cmu.edu, mpruden@kennesaw.edu

Abstract

Online communities of involuntary celibates (incels) are a prominent source of misogynist hate speech. In this paper, we use quantitative text and network analysis approaches to examine how identity groups are discussed on incels.is, the largest black-pilled incels forum. We find that this community produces a wide range of novel identity terms and, while terms for women are most common, mentions of other minoritized identities are increasing. An analysis of the associations made with identity groups suggests an essentialist ideology where physical appearance, as well as gender and racial hierarchies, determine human value. We discuss implications for research into automated misogynist hate speech detection.

1 Introduction

Warning: this paper contains content that is disturbing, offensive, and/or hateful.

Online communities of those calling themselves "involuntary celibates," (incels) are known for online misogynist hate speech and offline violence targeting women, including incidents of mass violence in Isla Vista, California, in 2014 and Toronto, Canada, in 2018, among others. Though some work in natural language processing (NLP) has focused on features of misogynist language in general (Anzovino et al., 2018; Samghabadi et al., 2020; Guest et al., 2021), online incel communities are known for significant lexical innovation (Farrell et al., 2020; Gothard, 2021). Training with data from incel forums would enable misogynist hate speech classifiers to identify the neologisms and novel ideological features of this dangerous form of online misogyny (Jaki et al., 2019).

In this paper, we provide hate speech researchers with a quantitative overview of trends and particularities of language in one of the largest misogynist incel communities, incels.is, which launched in 2017 following the r/incels ban from Reddit.

We focus this analysis on mentions of identities, which are key to automatically identifying hate speech (Uyheng and Carley, 2021) and a window into the ideologies of social movements (Benford and Snow, 2000). We ground this analysis in theoretical approaches that focus on how identities are *constructed* in interaction (Bucholtz and Hall, 2005; Burr and Dick, 2017) and investigate the following research questions:

RQ1. How frequently are different identities, including novel terms for identities, mentioned in incels.is discourse?

RQ2. How do identity mentions in incels.is discourse change over time?

RQ3. How are identity mentions used differently by central incels.is users?

RQ4. What textual associations are made with identity groups on incels.is?

To address these research questions, we first measure the distribution of identity term mentions using a large generic list of identity terms combined with community-specific identity terms surfaced from a word embedding-based approach. We confirm the most frequent identity mentions in this data are for women, with almost one-fourth of these being derogatory community-specific neologisms, such as "femoids." Mentions of gender are much higher than in a comparative white supremacist dataset, a similar commonly-used source of unlabeled hate speech (Simons and Skillicorn, 2020; Alatawi et al., 2021). We find increasing mentions of other minoritized identities, such as Black, LGBTQ+ and Jewish people on incels.is, suggesting a consolidation with broader far-right discourses. Users who are central to the network proportionally mention more of these other marginalized identities. From a quantitative analysis of the immediate contexts in which identity term mentions appear, a

pervasive hatred of women is clear, as well as reinforcement of stereotypes about other marginalized groups. The incel identity itself is often discussed with themes of victimhood and boundary-keeping for "true" versus "fake" incels.

Throughout our analysis, we find evidence of an essentialist, black-pilled ideological framework where physical appearance determines the value of individuals and groups. While rigid racial and gender hierarchies are not new (e.g., eugenics) and are often circulated in far-right discourse (Miller-Idriss, 2022), this incel community attaches many novel measurements to appearance related to these hierarchies, re-entrenching and extending them. We argue that to detect such a particular form of extremism, hate speech researchers must heed both the jargon and deeper ideologies of this movement.

2 Incels and Male Supremacism

Online misogynist incel communities are situated within a set of anti-feminist groups often termed the "manosphere." These groups include Men's Rights Activists, Pick Up Artists (PUAs), Men Going Their Own Way (MGTOW). Such groups are often associated with a "red pill" ideology, a *Matrix* film reference to seeing the hidden truth behind the world, in this far-right context a view that feminism has brainwashed and subordinated men (Ging, 2019). In addition, incels often refer to a "black pill," the idea that they are genetically predetermined to be incels and cannot improve their situation through work or self-improvement (Pruden, 2021). This leaves many black-pilled incels feeling that their only options are to cope, commit suicide, or commit mass violence (expressed in the common phrase, "cope, rope or go ER [Elliot Rodger, an incel mass shooter])." Among groups in the manosphere, incels are most associated with violent and high-profile events that demonstrate "extreme misogyny" (Ging, 2019).

Ribeiro et al. (2021) find that incel communities are both more extreme and more popular than older, more moderate male supremacist movements, while LaViolette and Hogan (2019) find more extreme manosphere movements contain essentialist, deterministic ideologies of identity, which we also find in incels.is. We find evidence of this reductionist and biologically essentialist worldview in associations with identities in incels.is.

Qualitative research on male supremacist extremism frequently examines Men's Rights Activists (Berger, 2018) and more moderate groups that are less niche and without the emergent vocabulary common in incel spaces. We find this tendency extends to the data sources in automated hate speech research and argue for the importance of attending to the particularly dangerous discourse of black-pilled incels such as those on incels.is.

Quantitative and computational studies of the manosphere often focus on the unique misogynist language use of these communities. Gothard (2021) and Jaki et al. (2019) surface incel jargon by comparing word frequencies in incel Reddit posts with subreddits and Wikipedia articles outside of the incel movement, while Farrell et al. (2020) find frequent incel terms not present in English dictionaries and expand their lexicon with a word embedding space. Such word frequency analysis, as well as hand-crafted lexicons, are often used to measure and study misogyny in the manosphere (Heritage et al., 2019; Farrell et al., 2019; Jaki et al., 2019). Pruden (2021) and Perry and DeDeo (2021) use topic modeling to characterize narratives and map out user trajectories on incels.is and r/theRedPill, respectively. Jaki et al. (2019) use word frequency analysis to study identity construction on a similar forum, incels.me, though their 6-month dataset only enables limited time-series analysis. In contrast, our work focuses on the use and contexts of generic and community-specific identity terms beyond a sole focus on misogyny, as well as how identity term use changes over time.

2.1 Automated misogyny detection

In early work on automated misogyny detection, Hewitt et al. (2016) and Waseem and Hovy (2016) developed small Twitter datasets annotated for sexism. Anzovino et al. (2018) proposed a keyword-based annotated dataset and taxonomy for misogyny detection on Twitter, with later shared NLP tasks (Fersini et al., 2018b,a; Basile et al., 2019; Bhattacharya et al., 2020). Data for these tasks came from Twitter posts and YouTube comments based on keywords, profile information, and YouTube video topics. Such data sources may capture mainstream misogyny but miss the unique linguistic characteristics of the incel movement.

Other annotated hate speech datasets have included data from manosphere subreddits, such as r/MensRights, r/MGTOW, r/incels, and r/TheRedPill, along with many other sources of online hate speech (Qian et al., 2019; Sap et al.,

2020; Mollas et al., 2020). Guest et al. (2021) propose a Reddit dataset annotated for misogyny by trained annotators, who would be more likely to understand community-specific jargon than crowdworkers with limited training. Though they include a variety of manosphere-related subreddits, absent from this dataset are banned black-pilled incel subreddits such as r/braincels, r/shortcels, and r/incels, the precursor of the more extreme incels.is.

3 Data

Our dataset contains 6,248,234 English-language public comments posted between the forum's creation in November 2017 and scraping in April 2021.[1] It includes forum and thread names, as well as the date of posting, user names and the comment's full text. However, it does not contain images, which is a limitation.

White supremacist dataset We compare identity mentions on incels.is to another common source of unlabeled hate speech: white supremacist texts. From a large, multi-domain, English-language white supremacist dataset (Yoder et al., 2023), we select posts from online forums in a similar time frame as the incels data, 2015-2019 (the latest year available in the white supremacist dataset). This subset includes 3,410,623 posts from Stormfront, Iron March, and 4chan /pol/ in threads with fascist and white supremacist topics or posted by users choosing white supremacist, Nazi, Confederate or fascist flags.

4 Methods

We take a quantitative approach to studying discursive identity construction (Bucholtz and Hall, 2005; Gee, 2011), borrowing a focus on in-group and out-group identity presentation from social identity theory (Tajfel, 1974; Seering et al., 2018). Specifically, we examine the use of identity terms and the immediate contexts in which they appear. A few mentions of an identity group may not represent attitudes of participants, but associations repeatedly made over the course of a 6 million-post corpus are more likely to capture widely shared beliefs (Stubbs, 2001).

4.1 Measuring the use of identity terms

We first find identity terms using a generic lexicon combined from multiple sources: the extensive list of English identity terms from the NetMapper software (Joseph et al., 2016; Carley et al., 2018), as well as identity terms frequently found in hate speech (Yoder et al., 2022) and terms for LGBTQ+ and neurodiverse identities found online (Yoder et al., 2020; Yoder, 2021). This combined lexicon totals 19,050 unique identity terms. Ignoring case, 7,244 were present in the incels.is dataset.

Grouping identity terms We aggregate identity terms referring to similar groups (such as *LGBTQ+ people*) and then further group those identities into broader demographic categories (such as *gender/sexuality*). To form these groupings, we adapt identity terms group labels used in hate speech research from Uyheng and Carley (2020) and Yoder et al. (2022).[2] Intersectional identity terms are counted for all groups indicated by the term, e.g., "white women" was counted for both "white" and "women."

Identity lexicon expansion To capture the neologisms that incel communities are known for (Jaki et al., 2019; Gothard, 2021), we expand our generic identity lexicon to nearest neighbors in word embedding space, a common approach (Demszky et al., 2019; Simons and Skillicorn, 2020; Lai et al., 2021). We trained a 300-dimension word2vec model (Mikolov et al., 2013) over our data and manually examined terms appearing at least 1,000 times among the top 30 nearest neighbors by cosine distances to a) the 30 most frequent generic identity terms or b) the mean of identity term embeddings in an identity group. This resulted in 84 new terms, the most frequent of which are in Table 1.

Varieties of "incels" It is common in incel discourse to refer to different types of incels with terms including a "cel" suffix (Gothard, 2021). For example, "tallcels" refers to tall incels and the racist terms "currycels" and "ricecels" refer to South Asian and East Asian incels, respectively. Excluding usernames, over 1500 unique words used in our incels.is dataset contained the string "cel," many of which referred to varieties of incels. We examined the 100 most frequent words containing "cel"
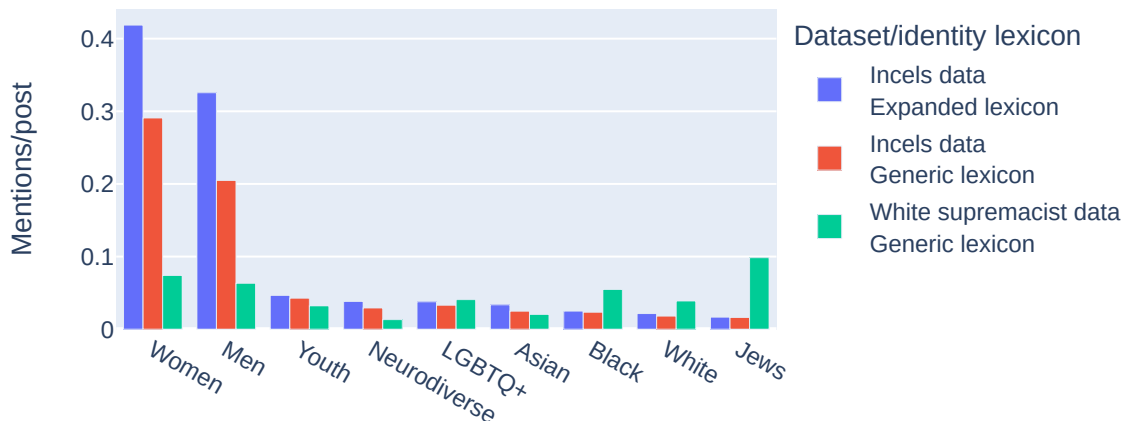

Figure 1: Identity group mention frequencies.

Community-specific identity terms
foids, chads, manlets, stacies, boyo, femoids, ethnics, chadlites, roasties, holes, betabux, landwhales, waifus, jbs, chicks, noodlewhores, soyboy, br0, aspie, betas, thots, traps, beckies, m8, boi

"Cel" variants
truecels, fakecels, volcels, greycels, escortcelling, gymcelling, ricecels, mentalcels, currycels, fatcels, femcels, whitecels, framecels, youngcels, oldcels, blackcels, ethnicels, brocels, itcels, incelistan, nearcels, tallcels, shortcels, locationcels, bluecels

Table 1: Most frequent 25 novel identity and "cel" terms found in the incels.is dataset. Plural and singular mentions are combined, as are "-ing" terms with their roots.

and grouped words that referred to incel variants, except those referring to "fake" incels, within the incels identity group for further analysis.

Central forum users We also analyze how forum leaders (prototypical incels) use identity terms. To find such leaders based on network structure, we construct a undirected graph where nodes are users and edges are weighted by the number of shared threads (out of 154,049 threads) between them. This graph contains 6819 users and 3,889,054 links. We operationalize central users as the top 5% ranked by eigenvector centrality, which measures if users share threads with other highly-connected users. These central users had a roughly similar number of posts as the rest of the users combined.

4.2 Associations with identity terms

Beyond the occurrence of identity term mentions, we analyze associations made with identities in their immediate contexts. Specifically, we extract *actions* taken by or to these groups, as well as *attributes* associated with them, a simple approach to analyzing the presentation of entities in discourse (Bamman et al., 2013, 2014; Yoder, 2021). For actions, we extract verbs where an identity term is the subject or object from a dependency parse. Attributes are adjectives and appositives whose head word is an identity term.

We surface the actions and attributes most distinctively associated with each identity group with PMI[3] (Daille, 1994; Role and Nadif, 2011), a variant of pointwise mutual information that lowers the ranking of low-frequency terms.

5 Results

5.1 Distribution of identity mentions (RQ1)

Prevalence of the most popular identity group mentions in our incels.is dataset is seen in Figure 1. Expanding the generic identity list with context-specific identity terms dramatically increases the detection of mentions of all identity groups, especially women, men, and neurodiverse people[3]. Adding these context-specific identity terms increases the total number of mentions identified from 6.46 million to 8.91 million–a jump of 37.8%–

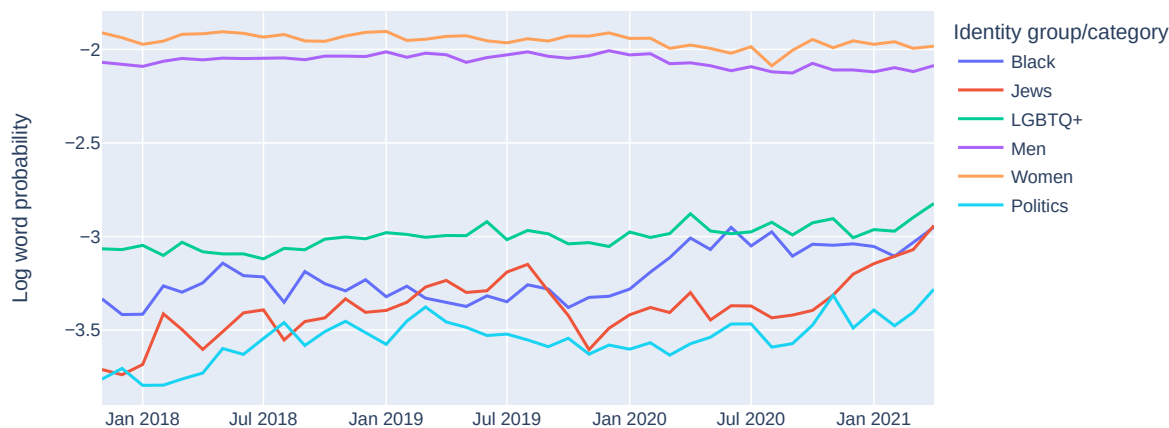[3]Many incels self-identify on the autism spectrum.

4

Figure 2: Selected identity group and category mentions over time in the incels.is dataset. Mentions of other identity groups remain steady.

demonstrating how common identity term innovations are in this community.

Also visible in Figure 1 is a comparison of identity group mention frequency with another common source of hate speech, white supremacist data. Mentions of women and men are much more frequent in the incel data than the white supremacist data, surpassing 0.4 mentions/post for women. Mentions of racial identities and Jewish people are more commonly found in the white supremacist data. This confirms that discourse from incel communities can be a useful source of misogynist text, especially after recognizing the lexical innovations referring to women and others.

5.2 Identity mentions over time (RQ2)

Figure 2 displays the prevalence of identity group mentions in this forum over time, binned every month during the dataset range and identified with the expanded lexicon. To control for any systematic changes in post word count over time, we present log word probability (the logarithm of identity group mention counts normalized by total word count).

Though mentions of women and men are most frequent across the data range, they stay steady or slightly decrease over time. There is a steady rise, however, in mentions of LGBTQ+ identities. Except for a decrease in the latter half of 2019, mentions of Jewish people also steadily rise. There is a significant rise in mentions of Black people in 2020, reaching a peak in June (during the anti-racist uprising against police brutality) and remaining at
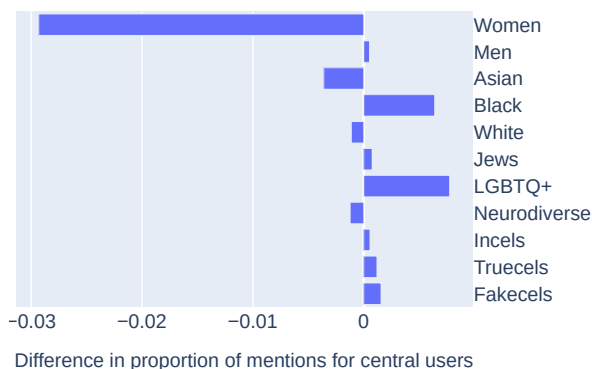

Figure 3: Absolute difference between the proportion of identity mentions used by top 5% central users in the shared thread network for each identity group and the proportion of mentions used by the rest of the users.

an increased rate through 2020 and 2021. Mentions of political identities also rise.

5.3 Central users' use of identity terms (RQ3)

Figure 3 shows the absolute difference in proportion of identity term mentions for the top 5% of users ranked by eigenvector centrality, compared to the rest. Proportionally, central users are less likely to mention identity terms for women, but are more likely to mention Black, LGBTQ+, and Jewish people. A concern for incel authenticity is also reflected in this central group's increased use of "truecels" and "fakecels" compared to other users.

Identity		Top PMI^3 terms
Women	*Attr*	white, old, single, fat, young, hot, fucking, ugly, cute, other, ethnic, average
	Act$_S$	want, get, love, care, go, hate, think, like, fuck, say, give, look, find, wants
	Act$_O$	get, fuck, hate, fucking, find, getting, having, attracted, see, want, fucked
Men	*Attr*	ugly, white, other, good, looking, average, tall, black, nice, short, chad, young
	Act$_S$	get, looks, go, need, look, think, want, got, mogs, fuck, going, become, gets
	Act$_O$	over, see, know, want, fuck, hate, fucking, love, seen, get, against, laid, date
Asian	*Attr*	south, central, >, east, average, half, ugly, other, skinned, northern, full
	Act$_S$	look, hate, get, cope, worship, go, need, mog, tend, eat, seem, make, want, take
	Act$_O$	learning, learn, speak, hate, seen, see, mog, killed, against, over, above, know
Black	*Attr*	north, real, west, other, fucking, man, stupid, dumb, ugly, dark, average
	Act$_S$	get, got, commit, slay, aspire, look, developed, gon, need, go, tend, fuck, run
	Act$_O$	free, hate, fuck, see, against, say, around, date, prefer, fucking, kill, sand
White	*Attr*	southern, northern, non, eastern, western, pure, white, nordic, other, average
	Act$_S$	go, get, want, mog, look, hate, invented, need, going, tend, voted, did, age
	Act$_O$	worship, hate, prefer, against, want, date, after, over, see, sought, towards
Jews	*Attr*	orthodox, fucking, rich, religious, secular, international, anglo, elite, greedy
	Act$_S$	control, did, created, want, pushing, won, own, pushed, made, win, took, push
	Act$_O$	hate, blame, ashkenazi, against, because, blaming, gas, kill, hated, hating
LGBTQ+	*Attr*	fucking, it, bluepilled, moral, low, normal, others, larping, stupid, ill, little
	Act$_S$	get, exist, go, say, need, think, fuck, try, deserve, trying, look, make, want
	Act$_O$	hate, fucking, coping, fuck, shut, ban, turning, kill, banned, larping, go
Neurodiverse	*Attr*	social, severe, functioning, fucking, extreme, crippling, sentence, complete
	Act$_S$	exist, makes, worse, causes, affect, get, comes, means, make, sucks, goes, go
	Act$_O$	because, due, diagnosed, cure, fucking, having, caused, cause, causes
Incels	*Attr*	fellow, other, blackpilled, true, real, actual, white, ugly, bluepilled, blackpill
	Act$_S$	get, exist, means, need, ascend, go, cope, know, want, say, become, going, look
	Act$_O$	help, against, hate, coping, create, see, creating, hates, bullying, die, ok, bullied

Table 2: Actions and attributes associated with identity group terms in incels.is dataset. *Attr* refers to attributes, while *Act$_S$* are actions for which the identity is a subject and *Act$_O$* are actions for which the identity is an object.

5.4 Associations with identities (RQ4)

Terms commonly associated with identity groups are presented in Table 2. Across groups, we find that the most frequent attributes relate to physical features ("ugly," "short," etc.). The use of these descriptors suggest hierarchies based on appearance, race and gender. This focus on physical appearance is apparent in top terms used to describe women, including "young," "fat," and "hot." Example uses show the hierarchies of appearance that incels apply to women: "some can't tell a beta female apart from a hot whore and so lump all types of the female sub species together."[4] Actions for which women are subjects suggest incels' speculation about what women "want," "love," or "hate." Common actions for which women are grammatical objects include "fuck" and "hate"– evidence of the forum's misogyny, casting women as things to be "attracted" or controlled.

Men are also discussed with an emphasis on physical appearance, as well as domination. "Ugly," "looking," "average," and "tall" are all top attributions for men. Top actions include "get," "need," and "mogs," an incel neologism meaning "dominate". An example post that reinforces gender hierarchies reads, "men literally mog femoids across the board, yet the foids whine about it."

Race is relevant in discussions of gendered identities. "White," for example, is a top descriptor for both women and men, as is "black" for men. Top terms suggest a negotiated association of superiority with whiteness. White people are the grammatical objects of "worship" but also of "hate." "Pure" and "nordic," common white supremacist descriptors, are distinctive attributes used for white people.

In contrast, Asian people are cast as subjects of actions like "worship" and "cope."

We find a range of common stereotypes for minoritized identities, particularly conspiratorial antisemitic tropes, as evidenced by terms suggesting a global Jewish conspiracy (e.g., "elite" and "control") and derogatory associations with the Holocaust ("gas"). One example post reads, "the endgame is an global Jewish Communist dictatorship," while another mixes antisemitic conspiracy theories with anti-feminism: "feminism is a subversive Jewish movement designed to ruin us."

LGBTQ+ characterizations are negative and associated with inauthenticity (e.g., "larping," or live action role playing). For example, one posts reads, "Lesbians don't exist. They're just bisexual foids who like women but still can't resist Chad."

Violence is associated with Black people ("commit"), for example in one post that reads, "I don't hate blacks because they're ugly, I hate them because no matter where they are they commit crime."

In-group identity associations Victimhood, race and authenticity are common themes associated with identity mentions of incels themselves and "incel" variants on the platform.

The most frequent lexical variations containing the "cel" suffix are in Table 1. Top terms relate to authenticity, including "fakecel," "truecel" and "volcel," ("voluntary celibate"), a focus that has also been observed in incel subreddits (Gothard, 2021). Platform affordances highlight distinctions between frequent and non-frequent posters: "greycels" who have posted less than 500 times have a gray-colored username and are often deemed inauthentic. Variants related to race ("whitecels," "ethnicels") are also frequent, suggesting the importance of race in incel self-classification (Jaki et al., 2019; Farrell et al., 2020). Also visible in these "cel" variations are a set of categories based on the familar theme of physical appearance (e.g., "fatcels" and "youngcels"). "Femcels," or female incels, are frequently mentioned, usually derided as outside the inherent masculinity of inceldom.

Incels are cast as merely "existing" or "coping," (Table 2) while others "hate" them or are "against" them. This victimhood includes common masculine tropes, such as a supposed inability to control themselves. From one post: "we can't control what we want, devaluation of women is a coping mechanism for not being able to elicit a biological response in them." Common far-right narratives of victimhood at the hands of corporations, Jewish people, and the media are also present: "Jews and the media hate incels, and the gaming industry is full of SJWs [social justice warriors]."

Race is also important–and controversial–in associations made with incels. "White" is a top incels attribute on the forum, and both "ethnic" and "white" are associated with truecels (see Table 3 in Appendix A for top terms related to truecels and fakecels on the forum). There is controversy over which races occupy what positions in an assumed hierarchy, often centering around the "just be white" (JBW) theory that white men have access to sexual relationships with women of all races. Some posts support this theory, e.g., "being white is a +3 when it comes to noodles [Asian women], so a 4/10 white is better than a 6/10 ethnic." Others challenge this notion: "a brown man with a chiseled face will mog a white incel everywhere." Still others echo the white supremacist Great Replacement Theory, blaming JBW as a way for incels of color to "get whitecels out so sh**skins can take over."

"Real" and "true" are top attributes associated with talk about incels, echoing a focus on authenticity in the top "cel" variants. This boundary-keeping is also visible in the words associated with fakecels (e.g., "detected" and "ban"). Authentic incels are victims of women's hatred ("if women aren't trying to kill you, you're not a true incel"), post a lot ("graycels are a joke with their tiny post counts") are unattractive ("I'm an incel, of course she said no to my hideous face"), have no female friends ("what true incel has a female friend? stupid newf*g") and do not date ("normie spotted. real incels are doing this all weekend and have no dates"). They also are unable to "ascend" (i.e., have sex and leave inceldom), and are "mogged" by others. Jaki et al. (2019) found similar themes in an earlier incel dataset.

6 Discussion

Across our quantitative analysis of the distribution and associations made with identity terms, we see evidence of an ideology where physical appearance determines human value, as has been found with prior work on incels (Maxwell et al., 2020; Baele et al., 2021; Pruden, 2021). This ideology essentializes social constructs, such as race and gender, as biological physical features impacting desirability, with controversy over the role of race.

We find strong evidence for gender as a cen-

tral focus of incel discussion; mentions of men and women far surpass the number of mentions of any other identity. We find that this community commonly uses novel identity terms that may not appear in generic lists, including many derogatory terms for women ("foids," "landwhales").

Increases in mentions of LGBTQ+, political, Jewish, and Black identities, often with stereotypes and conspiracy theories, could suggest this community has incorporated broader far-right trends. An increasing politicization is reflected in this example post: "we don't need society to completely accept the incel ideology, we just need to masquerade as normies and keep bashing women, jews and gays." Our evidence from text analysis supports the common user movement that Mamié et al. (2021) found from manosphere content to alt-right content on YouTube and Reddit. We find that many associations on incels.is reinforce stereotypes such as LGBTQ+ identities being fake, Black people being criminals, and antisemitic conspiracy theories. Users who are central in the forum's shared network devote more identity mentions, proportionally, to Black, LGBTQ+, and Jewish people compared to average users, suggesting that leaders on the platform play a role in broadening the discussion to include mentions of marginalized identity groups other than women. We also find more mentions of neurodiversity and mental health in this online community than in a dataset of white supremacist online content, which may be part of a victimhood narrative.

The overarching black-pilled ideology of physical appearance determining human worth also extends to talk about incels themselves on incels.is. Jaki et al. (2019) also find this "negative self-image" on a precursor forum, which is theorized by Nagle (2015) and Ging (2019). Though incels are presented as occupying the lowest status among men, we find fierce gate-keeping around who can claim the identity and its perceived victimhood. This echoes theoretical work by Kleinke and Bös (2015) finding that online communities often disparage less typical members along with out-groups.

Central users are active in discussions of authenticity. Such victimhood could lead "authentic" misogynist incels to pursue symbolic–or material–action against "fake" incels in the community but also against the perceived unjust system and the women they believe benefit from it.

Identities constructed in interaction are negoti-

ated (Bucholtz and Hall, 2010); we find contention around race in inceldom. "White" is associated with both true and fake incels on the platform, often in connection with the folk JBW theory that white men appeal to women of all races.

Implications for automated hate speech detection Central to many hate speech definitions is whether a text denigrates groups based on identity characteristics (Sellars, 2016; Sanguinetti et al., 2018; Poletto et al., 2021). Identity terms are, thus, a major indicator and concern for hate speech detection. In our analysis of identity construction on incels.is, we confirm that mentions of men and women identity terms are much more frequent than in a similar source of unlabeled hate speech: white supremacist data. Incel texts, then, may be a good source of unlabeled or annotated data for misogyny detection. The dangerous black-pilled ideology in particular is missing from current misogyny datasets (Guest et al., 2021). Such data should be considered, but the broader issue is a need for subject matter expertise in building such datasets for automated hate speech detection. Experts should be consulted to know where to look for training data so that specific types of hateful movements, with lexical or other linguistic innovations, are not overlooked.

We find that almost 30% of identity mentions in our dataset involve community-specific neologisms, often derogatory terms against women. Training hate speech classifiers on data that does not include these terms hinders the ability to detect this substantial source of contemporary online misogyny.

Our analysis also draws attention to the ideological associations being made with identities in this discourse space. We find problematic stereotypes against not only women, but also LGBTQ+, Black, and Jewish people. Thus, incel text data is not only a source of misogyny, but also reflects broader trends related to the mainstreaming of far-right beliefs. Particularly pernicious is a black-pilled ideology that physical appearance determines human value, a reinforcement and extension of essentialized gender and racial hierarchies. Hence, fatphobia, homophobia, ableism, and racism are all wrapped up in misogynist incel content. Automatically detecting this broader ideology may be unattainable or extremely difficult with machine learning techniques, but we emphasize practitioners and researchers should be aware of this ideol-

ogy. A narrow focus on hate against women from these communities will miss these important–and increasing–trends toward politicization and hate against other groups.

7 Conclusion and Future Work

The incel movement and the collective identity around it is a relatively new expression of male supremacism. In this paper, we use quantitative text and network analysis techniques to investigate how identities are constructed in discourse on one of the largest incel forums. We study the identity group mention frequency over time, as well as actions and attributes associated with them.

We find that talk about women and men dominates identity mentions on this forum, though mentions of marginalized identities commonly targeted by far-right groups increase from 2017-2021, appearing in textual contexts that propagate stereotypes. Many of these mentions use novel, community-specific identity terms that would be missed with generic lists of identities or hate speech training data from other contexts. Future work could systematically evaluate the ability of existing hate speech classifiers to handle this jargon, as well as the particularly dangerous black-pilled ideology.

This ideology is apparent in discussions of identities, including in-group ones, that reinforce rigid physical hierarchies based on attractiveness, gender, and race. We find race is a site of contention in discussions of who are "true" incels. Gatekeeping around incel authenticity is common.

Negotiation around race and inceldom, as well as intersectional racism and misogyny in incel forums would be a fruitful avenue for future work. This dataset could also be compared with other incel discussions, such as incels.me and earlier banned subreddits r/incels and r/braincels. The role of platform affordances and informal mentorship on the platform could be further investigated, as Perry and DeDeo (2021) mapped different user pathways in and out of r/TheRedPill. Further network analysis could reveal how the behaviors we identify, including a rise in mentions of marginalized and political identities, were spread in this community and why.

Limitations

Our approaches largely focus on explicit mentions of identity terms. This does not capture whether the identity term is the target of hate speech, which would require further analysis. This approach also does not capture attitudes held toward high-profile members of those groups, which play a role in circulating associations with identities (such as personal attacks on women in gaming or the use of "George Soros" as shorthand for antisemitic conspiracy theories). Future work may try to capture and measure these attitudes.

Incels.is is a large, popular forum for black-pilled incel discourse, which has a unique and extreme ideology that we argue is under-represented in current hate speech datasets. However, our analysis is limited to this forum, and the trends we identify may not apply to more moderate incel discourse (e.g., r/IncelsWithoutHate) or related online male supremacist movements, such as MGTOW and PUAs. Though these communities are known to have related, but distinct jargon (Farrell et al., 2020), we emphasize that researchers should recognize these lexical innovations in their annotations for hate speech and include a variety of these communities in training datasets for misogyny.

Ethics Statement

The Association of Internet Researchers (AoIR) acknowledges internet research is complex, dynamic and often involves many gray areas– specifically related to what constitutes human subjects, private versus public spaces and data versus persons (Markham and Buchanan, 2012). For this reason, the AoIR guidance recommends an inductive, ongoing and context-specific approach to ethics throughout the research process. At all stages, this involves being mindful of the vulnerability of the community under study and taking efforts to protect them where appropriate, while balancing their rights with social benefits and the researcher's right to conduct research.

Following this guidance, we subscribe to a utilitarian philosophy where we focus on doing the greatest good for the greatest number of people. In the case of black-pilled incels, we believe the necessity to better understand this potentially dangerous group outweighs the possible damage to forum members. For this reason, in addition to the AoIR guidance outlined above, we have followed some commonly accepted standards to protect participants and refrain from amplifying misogynist voices.

Data was collected only from publicly available online message boards and no private or identifiable information has been included in this manuscript.

We are not publishing user names, though we did observe them in our analysis of central users. We also did not subscribe to any channels or recirculate any content to ensure our work does not contribute to the monetization of the forum or associated accounts. Following the WOAH recommendation, we paraphrase posts to retain key aspects while protecting users' privacy.

References

Hind S. Alatawi, Areej M. Alhothali, and Kawthar M. Moria. 2021. Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding with Deep Learning and BERT. *IEEE Access*, 9:106363–106374.

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. Springer International Publishing.

Stephane J. Baele, Lewys Brace, and Travis G. Coan. 2021. From "Incel" to "Saint": Analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and Political Violence*, 33(8):1667–1691.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning Latent Personas of Film Characters. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 352–361.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian Mixed Effects Model of Literary Character. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 370–379.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63.

Robert D. Benford and David A. Snow. 2000. Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 26(1):611–639.

J. M. Berger. 2018. *Extremism*. MIT Press.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, and Yogesh Dawer. 2020. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Mary Bucholtz and Kira Hall. 2010. Locating Identity in Language. In Carmen Llamas and Dominic Watt, editors, *Language and Identities*, pages 18–28. Edinburgh University Press, Edinburgh.

Viv Burr and Penny Dick. 2017. Social Constructionism. In Brendan Gough, editor, *The Palgrave Handbook of Critical Social Psychology*, pages 59–80. Palgrave Macmillan UK, London.

L. Richard Carley, Jeff Reminga, and Kathleen M. Carley. 2018. ORA & NetMapper. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, volume 3. Springer.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. Thesis, Paris Diderot University.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2970–3005.

Tracie Farrell, Oscar Araque, Miriam Fernandez, and Harith Alani. 2020. On the use of Jargon and Word Embeddings to Explore Subculture within the Reddit's Manosphere. In *12th ACM Conference on Web Science*, pages 221–230, New York, NY, USA. Association for Computing Machinery.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96, New York, NY, USA. Association for Computing Machinery.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, Turin, Italy.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN 2018*, pages 214–228, Seville, Spain.

James Paul Gee. 2011. *An Introduction to Discourse Analysis: Theory and Method*. Routledge, New York.

Debbie Ging. 2019. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, 22(4):638–657.

Kelly Caroline Gothard. 2021. The Incel Lexicon: Deciphering the Emergent Cryptolect of a Global Misogynistic Community. Master's thesis, The University of Vermont and State Agricultural College, Vermont, United States.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Frazer Heritage, Veronika Koller, Alexandra Krendel, and Abi Hawtin. 2019. MANTRaP:A Corpus Approach to Researching Gender in Online Misogynist Communities. In *12th BAAL LGaS SIG*.

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*, pages 333–335, New York, NY, USA. Association for Computing Machinery.

Sylvia Jaki, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. 2019. Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2):240–268.

Kenneth Joseph, Wei Wei, Matthew Benigni, and Kathleen M. Carley. 2016. A social-event based approach to sentiment analysis of identities and behaviors in text. *The Journal of Mathematical Sociology*, 40(3):137–166.

Sonja Kleinke and Birte Bös. 2015. Intergroup rudeness and the metapragmatics of its negotiation in online discussion fora. *Pragmatics*, 25(1):47–71.

Mirko Lai, Marco Antonio Stranisci, Cristina Bosco, Rossana Damiano, and Viviana Patti. 2021. HaMor at the Profiling Hate Speech Spreaders on Twitter Notebook for PAN at CLEF 2021. Technical report.

Jack LaViolette and Bernie Hogan. 2019. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on Reddit. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, pages 323–334.

Robin Mamié, Manoel Horta Ribeiro, and Robert West. 2021. Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 139–147, New York, NY, USA. Association for Computing Machinery.

Annette Markham and Elizabeth Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Technical report.

December Maxwell, Sarah R. Robinson, Jessica R. Williams, and Craig Keaton. 2020. "A Short Story of a Lonely Guy": A Qualitative Thematic Analysis of Involuntary Celibacy Using Reddit. *Sexuality & Culture*, 24(6):1852–1874.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Cynthia Miller-Idriss. 2022. *Hate in the Homeland: The New Global Far Right*. Princeton University Press.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: an Online Hate Speech Detection Dataset. ArXiv: 2006.08328.

Angela Nagle. 2015. *An investigation into contemporary online anti-feminist movements*. Ph.D. Thesis, Dublin City University, Dublin, Ireland.

Chloe Perry and Simon DeDeo. 2021. The Cognitive Science of Extremist Ideologies Online. ArXiv:2110.00626.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. In *Language Resources and Evaluation*, volume 55, pages 477–523. Springer Science and Business Media.

Meredith L. Pruden. 2021. *"Maintaining Frame" in the Incelosphere: Mapping the Discourses, Representations and Geographies of Involuntary Celibates Online*. Ph.D. Thesis, Georgia State University, Atlanta, Georgia, USA.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4754–4763.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The Evolution of the Manosphere across the Web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.

11

François Role and Mohamed Nadif. 2011. Handling the Impact of Low Frequency Events on Co-occurrence based Measures of Word Similarity - A Case Study of Pointwise Mutual Information. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011)*.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 11–16.

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'18)*, pages 2798–2895.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of Social Identity Theory to Research and Design in Social Computing. In *Proceedings of the ACM Conference on Human-Computer Interaction*, volume 2 of *CSCW*, pages 1–33. Issue: January.

Andrew Sellars. 2016. Defining Hate Speech. Technical report, Berkman Klein Center.

B. Simons and D. B. Skillicorn. 2020. A Bootstrapped Model to Detect Abuse and Intent in White Supremacist Corpora. In *Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020*. Institute of Electrical and Electronics Engineers Inc.

Michael Stubbs. 2001. *Words and phrases: Corpus studies of lexical semantics*. Blackwell Publishers Oxford.

Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social Science Information*, 13(2):65–93.

Joshua Uyheng and Kathleen M. Carley. 2020. Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *Journal of Computational Social Science*, 3(2):445–468.

Joshua Uyheng and Kathleen M. Carley. 2021. An Identity-Based Framework for Generalizable Hate Speech Detection. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 121–130.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL-HLT 2016*, pages 88–93.

Matthew L. Williams, Pete Burnap, and Luke Sloan. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6):1149–1168.

Michael Miller Yoder. 2021. *Computational Models of Identity Presentation in Language*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Michael Miller Yoder, Ahmad Diab, David West Brown, and Kathleen M. Carley. 2023. A Weakly Supervised Classifier and Dataset of White Supremacist Language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Michael Miller Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Michael Miller Yoder, Qinlan Shen, Yansen Wang, Alex Coda, Yunseok Jang, Yale Song, Kapil Thadani, and Carolyn P. Rosé. 2020. Phans, Stans and Cishets: Self-Presentation Effects on Content Propagation in Tumblr. In *12th ACM Conference on Web Science (WebSci '20)*, pages 39–48.

A Additional Tables

Table 3 shows actions and attributes associated with "trucels" and "fakecels," common incel variants mentioned in incels.is.

Identity		Top PMI^3 terms
Truecels	*Attr*	biggest, truest, real, actual, giga, ultimate, legit, confirmed, certified, blackpilled, ugly, absolute, genuine, hope, ethnic, white, old, automatic, fellow, bigger, other
	Act$_S$	ascend, get, post, know, rise, knows, remain, go, looks, confirmed, relate, rot, understand, cope, use, tried, need, browse, make, suicide, spend, ldar, roped, take
	Act$_O$	pleasure, confirmed, banning, mog, rejected, bluepilled, help, laid, banned, born, save, doomed, over, seen, calling, see, die, excluded, bullying, dude, mock, mocking
Fakecels	*Attr*	fucking, larping, biggest, volcel, inb4, obvious, banned, known, defending, other, massive, fuck, tbh, confirmed, gtfo, potential, normie, likely, one, looking, users
	Act$_S$	detected, gtfo, confirmed, spotted, get, ascend, post, say, need, banned, try, posting, larping, fuck, smh, come, leave, coming, bragging, go, worry, ruining, invade, piss
	Act$_O$	ban, calling, banned, gtfo, weed, defending, expose, fucking, larping, spot, call, exposed, defends, exposing, smell, purged, banning, defend, confirmed, found

Table 3: Representative actions and attributes associated with truecels and fakecels identity group terms in the incels.is dataset. *Attr* refers to attributes, while *Act$_S$* are actions for which the identity group is a subject and *Act$_O$* are actions for which the identity group is an object.