SN Soc Sci (2021) 1:213
https://doi.org/10.1007/s43545-021-00220-8

ORIGINAL PAPER

Toxic language in online incel communities

Björn Pelzer[1] · **Lisa Kaati**[1] · **Katie Cohen**[1] · **Johan Fernquist**[1]

Received: 29 December 2020 / Accepted: 13 July 2021

Abstract

The internet-based *incel* subculture has evolved over the past decade on a number of different platforms. The subculture is known to be toxic and has become associated with several high-profile cases of lethal violence. In this paper, we study the level of toxic language and its targets on three large incel forums: *incels.co*, *lookism.net* and *looksmax.me*. These three forums are the most well-known and active online platforms where incels meet and discuss. Our results show that even though usage of toxic language is pervasive on all three forums, they exhibit significant differences in the composition of their toxicity. These differences correspond to different groups or philosophies within the incel communites.

Keywords Incels · Hate · Social media · Forums · Toxic language

Introduction

Several deadly acts of violence have been carried out in recent years by individuals who define themselves as *incels*. The term *incel* is a short form for *involuntary celibacy*, and it refers the members of an online community of digital meeting forums where the incel subculture has evolved. The incel culture is purely an internet phenomenon. There is no clear ideology that all incels agree upon, and there is no leader. What unites these individuals in the incel community is their lack of love and sex, and a feeling of never receiving recognition from women. These feelings and thoughts shape and fuel a community of

✉ Björn Pelzer
bjorn.pelzer@foi.se

Lisa Kaati
lisa.kaati@foi.se

Katie Cohen
katie.cohen@foi.se

Johan Fernquist
johan.fernquist@foi.se

[1] Swedish Defence Research Agency (FOI), Stockholm, Sweden

Published online: 09 August 2021

lonely and disappointed men who express hate towards women: They blame their involuntary celibacy on the women who reject them. Self-hatred and self-disgust are not uncommon among incels, and suicide is sometimes considered a clear-sighted way out of *inceldom*. For some individuals, the loneliness and hatred have become so overwhelming that it results in mass murder.

It is in digital environments where incels meet, form a community, and build a shared worldview. However, incels usually do not discuss their inceldom on large social media platforms like Facebook and Twitter. Rather, they meet in internet forums run by private individuals who are incels themselves, or who sympathize with them. The discussion forums serve as a kind of "safe space" for incels where they can make their own rules and speak their minds without being mocked, questioned and criticized. They are allowed to write hateful comments about women, whereas the moderators sanction criticism of incels or incel culture. The incel subculture is a part of the larger *manosphere* that includes a number of web-based misogynist movements focused on "men's issues" (Ribeiro et al. 2020) and who assert that feminism has gone too far and/or that men are actually the wrongfully oppressed gender (Ging 2019).

There is research indicating that the viewpoints expressed on the forum *incels.co* (at the time still named *incels.me*) exhibits some of the features that characterize extremist worldviews. These features include the notion of a strict and impenetrable hierarchical world order, as well as a clear division into "we" and "them," where cohesion within the group is strengthened by constant degradation of the other group (Baele et al. 2019).

Authorities are increasingly aware of incels. In January 2020, the Texas Department of Public Safety stated that there is a growing terror threat from offenders who identify themselves as incels (Texas Fusion Center 2020). The Swedish Security Police has also noted the terror threat from incels (Holmberg 2020). So far, a number of violent attacks have been committed by individuals identifying themselves as incels. One of the most well-known is offender is Elliot Rodger, a young man who killed six people in Southern California after posting a video online in which he stated that his violent actions were motivated by his experiences of social rejection, especially romantic rejection from women. Rodger later died by suicide and became seen as a martyr by many in the incel community (Glace et al. 2021). Another attack was done by self-identified incel in Toronto, Canada, who drove a van into a crowd of people, killing 10 (Beauchamp 2019). When it comes to suicide, sexual violence or harassment against women, there are no statistics that show the extent of connection to incel forums.

The digital incel environments are known to have a toxic discussion climate, featuring rape fantasies, suicide plans and glorification of mass murderers. While some studies have focused on investigating the level of hate and misogyny on internet forums that are a part of the greater manosphere, this paper focuses solely on digital communities for incels. We will not only measure the level of toxic language within incel communities, but we will also categorize the toxicity by its targets. The aim of our study is twofold: First, we want to measure the level of hate on three different digital platforms dedicated to incels. Secondly, we want

to analyze what the hate is directed at, so that we can categorize the hate and gain a better understanding of the subgroups within the incels.

To facilitate a discussion on the possible consequences of exposure to the toxicity in incel environments, we will also compile data about the typical activity lifespans of forum members, that is, over which period of time users tend to participate in the forums.

Overview

This paper is structured as follows:

In Sect. "Background" we introduce the incel subculture, and we review the current state of detecting toxic language in incel environments.

Section "Methods" describes the methodology used for this study, including our automated detector for toxic language, as well as the sample of incel forums that are the subject of our analysis.

In Sect. "Results" we present the measurements of the toxic language levels of the incel forums, and the distribution of hate/toxic language between the targeted groups. We also provide statistics on time periods over which users typically participate in incel forums and thus expose themselves to the toxicity.

In Sect. "Limitations of the study" we discuss the limitations of the study.

Finally, we discuss our findings in Sect. "Discussion", where we characterize the toxicity of incels, its motivations and consequences. We also provide recommendations for how society should respond to incels.

Background

Incels are not an organization or movement, nor is there any clear ideology that unites those who call themselves incels. Nevertheless, there are meeting places on the internet where a subculture with characteristic language, odd interests, and speculative theories has emerged among the participants. For many, participation in incel forums occurs during a limited phase in life. Others spend a lot of time over extended periods there, in an environment where the mood is characterized by depression and resignation alternating with rage, where mass murderers are portrayed as saints, and where an unfavorable appearance is considered a rational reason to commit suicide. In this section, we describe some of the language and theories that are unique for the incel culture.

Incels online

Online incel forums appear, disappear, and reappear under different names. One of the earliest discussion forums dedicated for incels was *PUAHate*, which was created in 2009 as a forum for men who felt that the seduction methods propagated by so-called pickup artists (PUA) did not work. Over time, the members of PUA-Hate developed the idea that appearance is crucial, and that all seduction methods

are meaningless to men who do not have the right appearance. PUAHate was shut down on May 24, 2014, the day after one of its members, Elliot Rodger, committed a mass murder. Two days after the shutdown, the forum *Sluthate* was started. Insufficient server capacity and internal conflicts led to dissatisfaction, and Sluthate was abandoned by its members when the new forum *lookism.net* emerged in June 2015. These three forums are jointly called PSL (Moonshot 2020). While two of the forums no longer exist, the abbreviation PSL lives on with several meanings in addition to the three forums, e.g. as a numerical unit for assessing a person's appearance ("*he is 4 PSL out of 10*") or as a designation of people who are or have been a member of one of the forums—"PSLers."

Since PUAhate was closed several forums for incels have emerged, sometimes temporarily or with changing names. In 2021 the three largest and most well-known incel forums are *Incels.is*,[1] *Lookism.net* and *Looksmax.org*.[2] During late 2019 and 2020 a number of new incel forums were created: *Blackpill*, *Non Cucks United*, *Looks Theory*, *You're not Alone*.

While incels have their own forums they also have created multiple communities on internet forums such as Reddit, 4Chan, 8Chan, and Tumblr (Ging 2019). Reddit allows users to create their own subforums, so called *subreddits*. The (now banned) subreddit */r/incels* used to be a popular incel environment with about 1.2 million posts. After the ban of /r/incels in November 2017 the subreddit */r/Braincels* gained popularity until it also was banned in October 2018. There are not only subreddits for incels to meet—there are also subreddits dedicated to critique of the incel ideology (Dynel 2020).

The incel worldview

The incel worldview is grounded in two beliefs: their understanding of society as a hierarchy where one's place is determined mostly by physical characteristics, and their identification of women as the primary culprit for this hierarchy (Hoffman et al. 2020). One of the essential concepts within the incel subculture is *lookism*. Lookism is discrimination based on appearance and the phenomenon that unattractive people are treated badly. Incels are convinced that their appearance is to blame for their involuntary celibacy and their difficulties in developing relations with women.

A popular explanation model discussed on the incel forums is that men are forced into involuntary celibacy due to feminism and women's liberation. Since modern women can have careers and be financially independent of their partners, they can also freely choose their partner. According to the 20/80 theory, when women are allowed to choose partners themselves, they tend to follow their genetically determined lookism and engage in so-called *hypergamy*, which means that they can select the most attractive men (Preston et al. 2021). The 20/80 theory claims that 80 percent of women will choose the "best-looking" 20% of men, to provide the best genes

[1] formerly *incels.co*.
[2] formerly *looksmax.me*.

to their offspring. On the other hand, men are considered to be less picky about appearance since they are genetically determined to maximize the number of sexual partners instead.

One of the most widespread phenomena within the incel culture is the *pill* metaphor. It has been appropriated from a scene in the 1999 film *The Matrix*, where the protagonist is offered a choice between two pills: The red pill reveals grim secret truths about the world, whereas the blue pill means continuing a life of delusion. In incel jargon, those who hold mainstream views are said to have taken the *blue pill* (or to be *bluepilled*), believing that personality is important to attract women, that society is generally fair, and that hard work pays off. The *red pill* represents the belief that genetics and looks determine all romantic success with women. Less handsome people can increase their romantic odds by improving their appearance (for example, through plastic surgery). A third metaphorical pill is the *black pill*: It shares the "redpilled" belief about the importance of genetics and looks but holds the fatalist view that improvements are impossible, and that a woman will never consider an incel as a partner (Preston et al. 2021). Incel forums frequently discuss whether only "blackpilled" people should be considered true incels. While the pill metaphor exists in many other subcultures, the black pill is unique to incels (Fernquist et al. 2020).

Detection of toxic language

Any cursory review of the discussions on the incel forums will reveal pervasive use of hateful and otherwise toxic language. To analyze this phenomenon in detail we first need to define our terms more clearly. Most people have an intuitive sense about what hateful language is, but there is no consensus on how to define it. Views are subjective, and levels of tolerance vary between individuals. A person may judge the same message differently depending on its context. The term *hate speech* is often used to cover different forms of expressions that spread, incite, promote or justify hatred, violence, and discrimination against a person or group of persons. While most definitions of hate speech share some common elements, alternative terminology such as *abusive language*, *toxic language*, and *dangerous speech* have been introduced to either broaden or to narrow the definition. A reason for the different terminologies is the need to differentiate illegal hate speech from legal expressions of hate or aggression toward certain groups, while still acknowledging that the latter may also be harmful. As we will see, verbal aggression in incel forums is often directed against groups that are not the traditional targets of hate speech. To characterize incel language we require a broader definition, and hence we will use the term *toxic language* to refer to blatantly aggressive and demeaning content, including hate speech.

Online toxic language detection

To get a deeper understanding of the toxicity in incel forums we need to survey large amounts of forum posts. As the forums have accumulated millions of posts, a

comprehensive manual analysis is not feasible, and automated methods are required. The automatic detection of toxic language, including hate speech and threats, can be regarded as a part of sentiment analysis and as such it has long been a subject of research in computational linguistics. The spread of the internet has also intensified commercial and political interest in such technology. Law enforcement agencies and civil rights groups investigate cases of online threats and hate speech, and social media like Facebook and Twitter are coming under increasing scrutiny and legal pressure to moderate the posts and discussions among their members, who number in the hundreds of millions. Arguably, though, such efforts tend to be more concerned with suppressing the symptoms of toxicity than with gaining insights into the underlying motivations.

In practice, the success remains modest, as the task of detecting toxic language is fraught with difficulties, beginning with the interpretation of the basic concept of toxicity itself. It is hard to derive any exact criteria which could be used to determine whether any given text is objectively toxic. Even human experts will often disagree on whether an author is hateful or maybe just angry, or even joking. Datasets intended for the training of automatic detectors based on machine learning therefore often include detailed annotation records, for example as with the set established in (Davidson et al. 2017). This allows the developer to decide on how to resolve disagreements between annotators before training, for example by majority vote.

Since humans will not always agree on toxicity, it is impossible for an automated approach to achieve any form of ideal detection results that would meet universal approval. This fundamental uncertainty of the task is further compounded by the general vagaries of natural language understanding: Computers lack any deeper comprehension of the meanings of statements, and they cannot reliably interpret subtle expressions, complex sentences, irony and sarcasm, jargon and misspellings. This is especially egregious with online toxicity, as users of internet forums like to be creative when expressing hate or threats, to appear witty or to avoid moderation or prosecution. This means that automated methods relying on a dictionary of toxic terms fall short as users constantly invent new terms, express hate or threats using individually harmless terms (e.g. "*I know where X lives, anyone up for a visit?*"), or use toxic terms in an ironic, joking or otherwise non-hateful manner. Automated systems will inevitably miss instances of toxic language and conversely also detect toxicity where none was intended by the author.

Another issue to consider when evaluating methods of toxicity detection is the widespread focus on hate speech, a predominantly legal category usually reserved for denigrating statements against groups or individuals based on their race, religion, sex or other innate attributes, and which thus only covers a subset of verbal expressions of hate. Systems and datasets for hate speech have limited usefulness when trying to characterize the unique features of incel toxicity.

State of the art

For the reasons outlined above, existing approaches to toxicity detection vary greatly in their methods and even in their definitions of hateful or toxic expressions. Comprehensive overviews of the field are provided in Schmidt and

Wiegand (2017) and Fortuna and Nunes (2018). While these attempts have been welcomed by many, others have expressed concerns about the utility of such solutions. It is, for example, argued that these solutions can be easily deceived, that they lack validity and reliability, and that they might reinforce existing societal biases (Sap et al. 2019).

In Malmasi and Zampieri (2017) machine learning is used to identify tweets which contain hate speech or profanity. Specifically, the authors employ a support vector machine on three different sets of features: character *n*-grams, word *n*-grams and word skip-grams. The results show that distinguishing profanity from hate speech is very challenging. Bag-of-words approaches tend to yield high recall, but at the same time they also lead to high rates of false positives, since the presence of offensive words can lead to the misclassification of texts as hate speech.

This phenomenon has already been described earlier (Kwok and Wang 2013), with the authors finding that tweets being classified as racist was mostly due to the presence of offensive words.

In ElSherief et al. (2018) it is noted that determining whether a text constitutes hate speech is more nuanced than *yes* or *no*. The authors distinguish between generalized hate speech (against groups) and directed hate speech (against individuals). They combine word lists derived from *Hatebase* (Hatebase 2020), manual curation and the *Perspective* API for toxicity detection (Jigsaw and Google 2020) to find tweets matching these categories and to identify differences in language use.

The difficulties of hate detection are in focus in Hosseini et al. (2017), where the authors show how to deceive the aforementioned Perspective API. The website of the API acknowledges that the system cannot recognize hate when it is expressed in ways the system has not encountered before.

One way to encounter the difficulties with hate speech detection is to develop domain specific hate speech detecting algorithms, i.e. algorithms that are trained to detect a specific type of hate speech. Several approaches using specialized algorithms have been developed, including algorithms for detecting jihadist hate speech (Ashcroft et al. 2015; Kaati et al. 2015), anti-Muslim hate speech (Olteanu et al. 2018), anti-black hate speech (Kwok and Wang 2013), misogynistic hate speech (Frenda et al. 2019; Saha et al. 2018), and anti-immigrant hate speech (Ross et al. 2016).

The *Bag of Communities* approach (Chandrasekharan et al. 2017) uses machine learning on posts from different communities to train a classifier for abusive online behavior. The training samples are labeled as abusive or non-abusive posts based on in which online community they were posted, with the communities assumed to be either entirely abusive or entirely non-abusive. We find this premise to be questionable, and while the method achieves a good accuracy of 0.75, it is unclear whether it actually detects abusive language or rather similarity to broader classes of communities.

The Online Hate Index (OHI) (Anti-Defamation League (ADL), 2018) is a tool intended to detect and quantify hate speech in online environments. The OHI is trained on 9000 annotated comments from Reddit. While intended for comparisons between different environments, the OHI project is in an early phase, and at the time of this writing it has not been tested on any other data beyond its training set.

Turning towards the domain investigated in this article, several other research projects have studied toxicity in the context of incels or the manosphere in general.

The level of hate speech on the forum incels.co is investigated in Jaki et al. (2019). The authors use a dictionary-based approach to identify posts with keywords that constitute misogyny and racism. They find that about 30% of the threads are misogynistic, and 3% are racist. About 5% of all messages contain one or more of the 10 offensive words they had listed.

The authors of Farrell et al. (2019) investigate misogyny on Reddit. The analysis is based on 6 million posts published between 2011 and December 2018, using lexicons designed to capture specific misogynistic rhetoric. The results show that there are increasing patterns of misogynistic content and users as well as violent attitudes on some of the Reddit subforums (*subreddits*) that can be seen as part of the manosphere.

The incel community on YouTube is studied in Papadamou et al. (2020). The authors report an increase in incel-related activity on YouTube, and that incel-related videos attract a substantially larger number of negative comments compared to other videos.

Glace et al. (2021) performed a content analysis of 400 top-rated posts from /r/Braincels and found that hostile sexism was an inductive theme in the conversations, including shaming women's appearances, degrading women for engaging in sexual activity, suggesting that women make false claims of rape and misogyny, asserting that women's only value is sexuality, and dehumanization of women. Another inductive theme identified by the authors was suicidality/meaninglessness. Maxwell et al. (2020) studied the sub-Reddit r/Braincels to get insights to shared experiences, sentiments and expression of self-defined incels. They found that the incel community on /r/Braincels demonstrated many deep-seated negative beliefs about women and the role of women in society and that women were depicted as vile, toxic, evil creatures who manipulate and destroy men (Maxwell et al. 2020).

In Ribeiro et al. (2020) a study of the internet based manosphere is presented. A total of 28.8 million posts from 6 forums and 51 subreddits are analyzed regarding the spread of toxicity and misogyny. The authors measure the toxicity of the forums using the aforementioned Perspective API, and for misogyny they use a misogyny lexicon. Their analysis indicates that some of the newer communities in the manosphere are more toxic and misogynistic than the older ones.

Our contribution here will be a deeper look into the toxicity among incels as the most notorious subculture within the manosphere. We quantify and categorize their toxicity in greater detail than earlier works, providing insights into their motivations and worldviews.

Methods

For our study we identified and retrieved suitable data in the form of posts from three large incel forums. This data was first analyzed automatically using a machine learning classifier that we had developed for the detection of toxic language. To achieve a more fine-grained characterization of the toxicity we then manually studied and

annotated a sample of those posts that had been flagged as toxic by the classifier, categorizing them by the targets of their toxicity.

Data

In this work we have analyzed forum posts from the three incel forums *incels.co*, *looksmax.me* and *lookism.net*.[3] These forums are some of the most well-known online environments where incels meet and discuss. All three are open forums (i.e. one can read posts anonymously, without registering), and they all refer to themselves as meeting places for incels. The three forums are described below. Statistics about the number of posts and registered members on the forums were retrieved in January 2020. Statistics about visitors have been obtained from the web analysis tool *SimilarWeb*,[4] which based its analysis on data obtained between October and December 2019.

The forum incels.co is the largest active digital environment for incels with around 11,000 registered members and 3.3 million posts. The forum was founded in November 2017. Incels.co is a forum for men only, and women are excluded according to the forum rules. The rules also forbid boasting about sexual or romantic experiences, and to publish pictures of "ugly" men with women, since such pictures are considered to give a false picture of reality. According to SimilarWeb, incels.co had more than 900,000 visits during the period October–December 2019. The forum has approximately 53,000 unique visitors each month. Almost 7000 members have written posts on the forum since it started in November 2017.

The forum looksmax.me is run by the same owner as incels.co. It is intended as a forum for men who want to discuss options to improve their appearance, with the goal of increasing their success with women. The forum does not allow any female members. Officially the members are not required to be incels, and anyone who is interested in appearance improvements is welcome. Unlike incels.co, the users of looksmax.me are allowed to post photos of themselves and to ask for advice and assessments ("rate me"). Incels.co refers its members to looksmax.me for this type of discussion. Looksmax.me was founded in August 2018. According to its own statement it has almost 1.4 million posts and 3400 members. During the period of October–December 2019, looksmax.me had just over 800,000 visits, and almost 39,000 unique visitors each month according to SimilarWeb's statistics. Since the forum started in August 2018, 2500 members have published posts in the forum.

The forum lookism.net was founded in June 2015 and is thus one of the oldest still active incel environments on the web. With over 10,000 members and 3.8 million posts, the forum is even bigger than incels.co, though no longer quite as active. Lookism.net includes both general discussions, appearance advice and methods to improve one's relationship status. Unlike incels.co the forum has no special

[3] Incel forums frequently change their web hosting providers, which can affect their domain names. We refer to the forums by the names they had at the time of our data retrieval in January 2020. Since then, *incels.co* has changed into *incels.is*, and *looksmax.me* has become *looksmax.org*.

[4] https://www.similarweb.com.

Table 1 Number of posts retrieved from each forum and timespan covered

Forum	Number of posts	From	Until
incels.co	3.3 million	November 2017	January 2020
lookism.net	3.8 million	June 2015	January 2020
looksmax.me	1.4 million	August 2018	January 2020

regulations on who may become a member. According to the user rules of the forum, discussions of illegal acts, such as rape, are prohibited, and so are racist comments. According to SimilarWeb's statistics, lookism.net had almost 400,000 visits between October and December 2019, with approximately 41,000 unique visitors per month. Since the forum started in June 2015, more than 10,400 members have published posts in the forum. Technical problems and hacking have reduced the popularity of lookism.net. In October 2019, the forum was vandalized and all posts were deleted. For several weeks the forum operated only sporadically and in a temporary manner. The recovery of deleted posts and full restoration did not begin until December 2019, and the forum did not preserve posts that had been published during the intermittent phase before the recovery. Therefore, there is a gap of approximately two months in the set of posts available on lookism.net. Given that posts are available from June 2015 onwards, we estimate this gap to have little impact on our data.

Data retrieval and preparation

The hierarchical structure of internet forums renders them impractical to analyze directly online. For example, it is not possible to retrieve a representative sample of posts directly from the respective server. Furthermore, forums differ in how they represent meta-data, for example the formatting of dates and how quotations are distinguished from other texts. This makes it difficult for a single analysis tool to process multiple forums. The first step in our analysis, was therefore, to download all the posts from the forums and to store them in a uniform format in a local database. For this purpose, we developed web spiders which traversed the forum hierarchies while downloading all of the posts available and standardizing their meta-data. We also stripped the posts of user signatures and quotations from other posts, to avoid data duplication and to ensure that the subsequent analysis would examine only the original text written by the respective author of each post.

The retrieval was performed during early January 2020, and our data thus consists of all public posts available at the time of retrieval, i.e. posts that had been published at any time before the retrieval and not been subsequently deleted. See Table 1 for a more detailed breakdown of the data per forum.

All the posts we downloaded are open to the public: No forum membership is required to read them, and large search engines like Google include the contents of these posts in their search index. We reason that these public posts carry implicit consent to being read.

For a comparison of incel forums to the "normal internet" we rely on data from Reddit, an online forum platform that is one of the most-visited websites. Reddit

users can freely create new subforums, and with over 130,000 active subforums Reddit has communities discussing virtually any topic of interest. Our dataset consists of 6.4 billion posts that we had retrieved earlier, covering the timespan from June 2005 until September 2019.

The datasets are still too large for a complete analysis of all posts. However, storing them in a local database enabled us to compile representative sample sets for all the forums. They also made it possible to compute statistics about user activity that are not provided by the forums themselves—these results are presented in Sect. "Activity and member lifespan".

Creating a toxicity classifier for incel forums

To investigate toxicity in incel forums we have trained a classifier based on the state of the art in toxicity detection and machine learning in general, while also drawing from our own earlier experiences with toxicity detection. In Isbister et al. (2018) we performed a real-world case study using dictionaries and CBOW-embeddings (Continuous Bag of Words) (Mikolov et al. 2013) to monitor online hate during the Swedish elections. The results were satisfactory from a pragmatic perspective, but at the time we could not evaluate accuracy and recall in a stringent manner. In another approach we combined dictionaries, natural language processing and automated reasoning to perform semantic analysis on forum posts and to identify those containing toxic expressions (Pelzer et al. 2018). The method cuts down on false positives and detects difficult occurrences, but overall recall is poor.

Later in Berglind et al. (2019) we used transfer-learning to train a toxicity classifier based on Google's BERT (Devlin et al. 2018). BERT is a language model, a neural network trained on thousands of books and Wikipedia; it has learnt the relations between words and thereby possesses a rudimentary understanding of language that forms a helpful base for subsequent training in more specific tasks such as toxicity detection. Language models are a promising approach to automated text analysis: Where a conventional machine learning model starts from a blank slate and then learns any language capabilities entirely from the annotated training samples, a language model already has a comprehensive linguistic foundation, and therefore it can be trained for specific tasks like toxicity detection using relatively few annotated samples. Accordingly, we fine-tuned BERT to detect toxic language by training it with about 20,000 samples from different hate speech and toxic language datasets, see Berglind et al. (2019) for details. The resulting model can therefore be said to employ an aggregated concept of toxic language, and its performance is in line with the state of the art, achieving an accuracy of 81 percent compared to a ground truth of averaged human assessments. This is the model that we also use as a classifier in the current analysis of toxic language among incels, and we employ it to measure the proportion of toxic posts—the *toxic language level*—of the three forums.

Due to the fundamental uncertainties of the task, we do not claim that our measurement of toxic language levels represents an exact measurement of the toxicity in each forum—there will be false positives as well as false negatives. However, by applying our uniform method across multiple forums, the results should provide

insights into how the forums compare to each other, and in particular to Reddit representing the "average internet."

Measuring the prevalence and distribution of toxic language

To measure the toxic language levels in the three different incel forums we used the model described in Sect. "Creating a toxicity classifier for incel forums". We measured the proportion of posts that contain some form of toxic language. For this we selected a random sample of 17,000 posts from each forum, ensuring a confidence level above 99% and a margin of error below one percent. Then we let our model analyze each post in the sample sets, flagging those it determined to contain toxic language. We did not account for the length of each text, nor for how often toxic language was used within one text and at what intensity. An analogous sampling and automated analysis were performed on the Reddit dataset. The resulting toxic language level measurements are presented in Sect. "Toxic language levels".

To get a deeper understanding of the toxicity on the incel forums, we conducted a manual analysis of posts that had been flagged as toxic. Our goal was to annotate the specific targets of the toxicity. For this we randomly selected 500 toxic posts from each incel forum. These were then annotated in two phases. The first phase served to derive the categories themselves: Out of the three sample sets we randomly selected 100 toxic posts per forum and examined them, taking notes of the respective targets of toxicity and their prevalence. We then grouped targets into categories, resulting in the following seven categories of toxicity:

- *Women*: Toxicity towards women is expressed in the form of derogatory words ("whore," "slut"), grossly negative comments about women's appearance, behavior and thinking, and in some instances calls for or fantasies about violence against women.
- *Society*: Hatred against the "system," and against "normal" people who are uncomprehending towards incels.
- *Incels*: Incels who claim that they hate incels as a group, or outsiders who registered on the forum and express their disgust with incels.
- *Self-hatred*: This category includes posts where the writer describes himself as abominable, idiotic, hopeless, etc.
- *Ethnicities*: Racist expressions, which may be aimed at others, but also at the author's own ethnicity.
- *Forum Users*: Toxic language aimed at another forum user, either directly in a conversation, or when talking to others about some user not directly participating in the discussion. This may be due to conflict and malice, but sometimes it is used as a "bitter pill," in order to reduce what is perceived as false hope.
- *Other*: Toxic comments that do not fit into any of the previous categories, e.g. hate directed at certain politicians, writers, films, books, etc. These target groups are individually too rare to form their own categories, and the toxicity often appears to be unrelated to incel ideas.

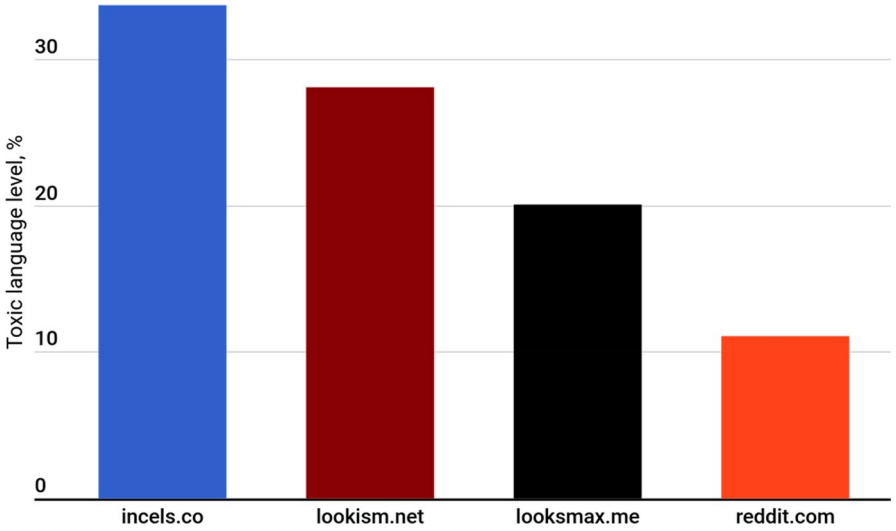

Fig. 1 The levels of toxicity in the three incel forums compared to Reddit

In the second phase we revised the annotations from the first phase to account for the seven identified categories, and we used the categories to annotate the remaining 400 toxic posts per forum. If a post fit into multiple categories, we counted it once for each. With 500 annotated posts per incel forum we have a confidence level of over 95% and a margin of error below 5% with respect to the subsets of posts marked as toxic by the classifier. The results are presented in Sect. "Toxic language levels".

To get a better perspective on how the toxic language levels may affect the audiences of incel forums, we compile statistics on how many months the members tend to participate in the incel forums. This data is presented in Sect. "Activity and member lifespan".

Results

In this section we present the results of our investigations as detailed in Sect. "Methods".

Toxic language levels

Our first measurement was performed using the machine learning classifier, measuring the toxic language level—the percentage of toxic posts in the three incel forums and in Reddit, the latter representing the "normal internet" in this study. The result is shown in Fig. 1. All three incel forums contain a significantly larger share of toxic posts than Reddit. There are also major differences between the different incel forums. With a toxic language level of 33.7% incels.co contains more than three

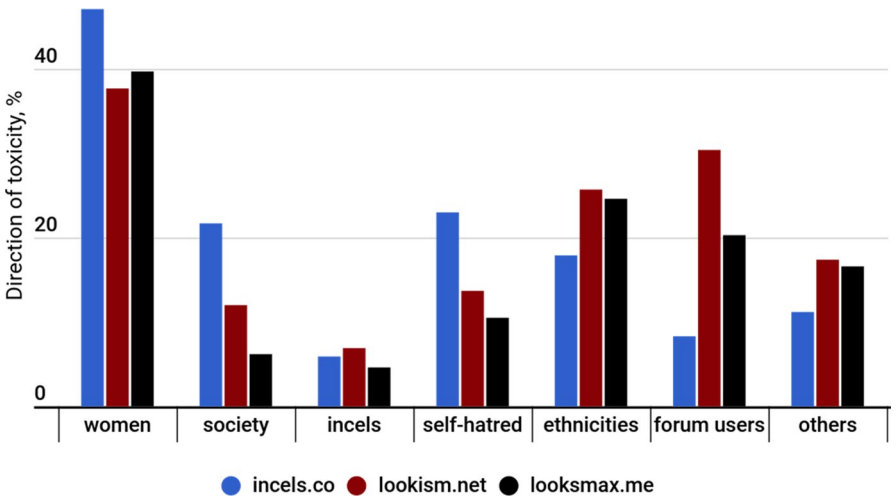

Fig. 2 Toxicity categories: the targets of toxicity in incel forums

times as many toxic posts as Reddit with its level of 11.1%. Lookism.net (28.1%) and looksmax.me (20.1%) are situated between these extremes, but both remain significantly more toxic than Reddit.

We then manually measured the distribution of toxic language over different target groups, and the result is presented in Fig. 2. Each bar represents the percentage of posts among the 500 toxic posts that contain the respective category of toxicity. Posts containing multiple categories of toxicity are counted in each category. Consequently, the sum of percentages for a forum may exceed 100%.

The results shows that the largest proportion of toxic comments in all three incel forums is directed towards women, ranging from 37.7% on lookism.net to 47.1% on incels.co. The hate against women manifests itself in the forum jargon which contains numerous derogatory and dehumanizing terms for women. For example, the terms *femoids* or *foids* (from "female humanoid," used for women in general), *roasties* (women who have had multiple sexual partners) and *landwhales* (obese women) are often found. Women are generally considered to be governed by their animalistic nature, and thus incapable of both independent decisions as well as deeper feelings. In addition to appearance, status and money, dominant behavior is considered a trait that makes men attractive to women (Baele et al. 2019). This is occasionally emphasized as a justification for beating, raping or humiliating women. Almost never are women described as persons that can be related to, rather they seem to be considered as a natural resource.

Toxicity against society differs more among the three different forums. It is most common on incels.co (21.7% but drops down to 6.2% on looksmax.me. A possible explanation lies in the different outlooks of the forums: incels.co is the most "blackpilled" of the three, focusing on hopelessness and incel ideology, while the other two put more emphasis on self-improvement. Incels.co and looksmax.me refer to each other for the respective discussions and try to keep the topics separate.

Toxicity against incels as a group is generally rare. Moderators do not allow generalized hate towards incels, and the users on incels.co are not allowed to disagree with incel views. This might be one of the reasons why the level of toxicity towards incels is quite low.

Self-hatred on the other hand is the second most common form of toxicity in incels.co, reaching 23%. Its distribution among the three forums is similar to that of toxicity against society. The reasons may be analogous, as forums with a focus on self-improvement are less likely to attract members who have abandoned all hope.

Conversely, toxic language against non-white ethnicities is most common on lookism.net (25.7%) and looksmax.me (24.6%). An interesting aspect is the nature of this toxicity. While the forums do contain much racist hatred, incels are nevertheless an ethnically heterogeneous group (Jaki et al. 2019). Many incels with an ethnicity other than "white" believe that women do not even look at them precisely because of their ethnicity. These so-called *blackcels*, *ricecels* or *currycels* speak of themselves in a way that expresses internalized racism (Pyke 2010), that is, they identify with an ideology that oppresses and degrades them. Some of them are engaged in *whitemaxxing*, which involves trying to tone down the ethnic features in their appearance. As such the racism on the incel forums is not exactly comparable to the one found on for example forums for white supremacists.

Toxicity against other forum users is most common on lookism.net at 30.4%, indeed the second most common category on that forum. On the converse, this category is relatively uncommon on incels.co at 8.3%. There are two factors contributing to this difference. For one, incels.co is likely the most ideologically homogeneous of the forums, with the policies banning members who disrespect incel views, and also banning certain topics that are likely to upset incels, for example bragging about romantic achievements. In contrast, both looks-focused forums are more open to other types of members, in particular lookism.net, and also more permissive towards such volatile topics, and thus they contain more potential for conflict between members. The other reason was mentioned in Sect. "Data": lookism.net and looksmax.me feature many threads where members ask for evaluation of their looks and advice on how to improve themselves. The responses often use a drastic language that triggers our automated toxicity detector, and this language would be highly inappropriate in other venues, but in incel forums it can be perceived as part of a realistic assessment, rather than a hateful insult.

The category *Other* summarizes all toxic expressions that do not fit into any of the more specific categories above, making up between 11.2% of the hateful sample set on incels.co to 17.4% on lookism.net. Like other online communities, incels discuss interests, news and everyday events without any direct relation to the main topic of their forum. Thus, the toxicity occurrences in this category, aimed at celebrities, movies, books, computer games, etc., would not be out of place in the off-topic sections of non-incel forums, although the language may be stronger than usual.

Overall, our measurements indicate that toxic language is significantly more prevalent in incel forums than in the "normal internet" as represented by Reddit. Our deeper manual sampling of toxic posts provides some insight into the composition of this toxicity. Hate against women is the dominating category in all three incel forums. In other categories we can observe a notable difference between the

"blackpilled" incels.co on one side and the more "redpilled" looks-focused lookism.net and looksmax.me on the other: The former exhibits more self-hatred and toxicity against society, whereas the latter show more toxic language directed against ethnicities and other forum users. Furthermore, toxic language against ethnicities and forum users is often not representative of hate in a conventional sense, as it may be aimed at the own ethnicity or intended and received as honest advice, although such usage still contributes to a highly toxic digital environment.

Activity and member lifespan

In the previous sections, we determined that the incel forums have a high prevalence of toxic language. To better estimate the effects of these environments on their members we need to know how long the members expose themselves to this toxicity. Towards this purpose we analyse the full forum datasets (see Sect. "Data retrieval and preparation"), grouping users by their activity lifespan, meaning the time from their first posting to their most recent one. Arguably, a higher activity lifespan indicates greater tolerance for incel toxicity, and possibly also more acceptance of incel ideology. As new members join continuously, we need to distinguish inactive users, who stopped participating after their most recent post, from active users whose most recent post occurred shortly before our retrieval and who may have continued participating. Hence in each lifespan cohort we count members as active only if their most recent postings occurred no more than six months before our data retrieval in January 2020. Users who stopped posting before this are regarded as inactive; we assume that these users have ended their activity on the forum. Figure 3 shows the lifespans for the three different forums.

On incels.co there is great variation in the activity of the members and many only stay active for a short while before no longer posting. User retention is generally low, as approximately half of the members stayed active for no more than one month, and indeed 70% of all members who ever posted are no longer active. The forum had attracted some 400 new members in the month before our retrieval. Around the one-year mark of activity the inactive users become a minority in their respective cohort, indicating that users who stay longer tend to become permanent members.

Lookism.net exhibits a similar pattern as incels.co: About half of the members were only active for up to one month. Overall user retention is lower, and only 9.7% of the members were active in the 6 months before our data retrieval. This however is skewed by lookism.net being the oldest of the three forums. In the month before our data retrieval lookism.net had garnered over 400 new users, comparable to incels.co. This is atypical, though, as the outages of lookism.net in late 2019 have resulted in there being no active users in the 2–5 months cohorts, and the 0–2 months groups likely reflect pent-up user influx.

The looksmax.me user activity ostensibly follows the familiar pattern with the 0–1 month cohort being the strongest, but 73.9% of the total member base are still active. Of course, this must be viewed in perspective, as looksmax.me is the youngest forum and thus has had less time for users to leave again. Nevertheless, active

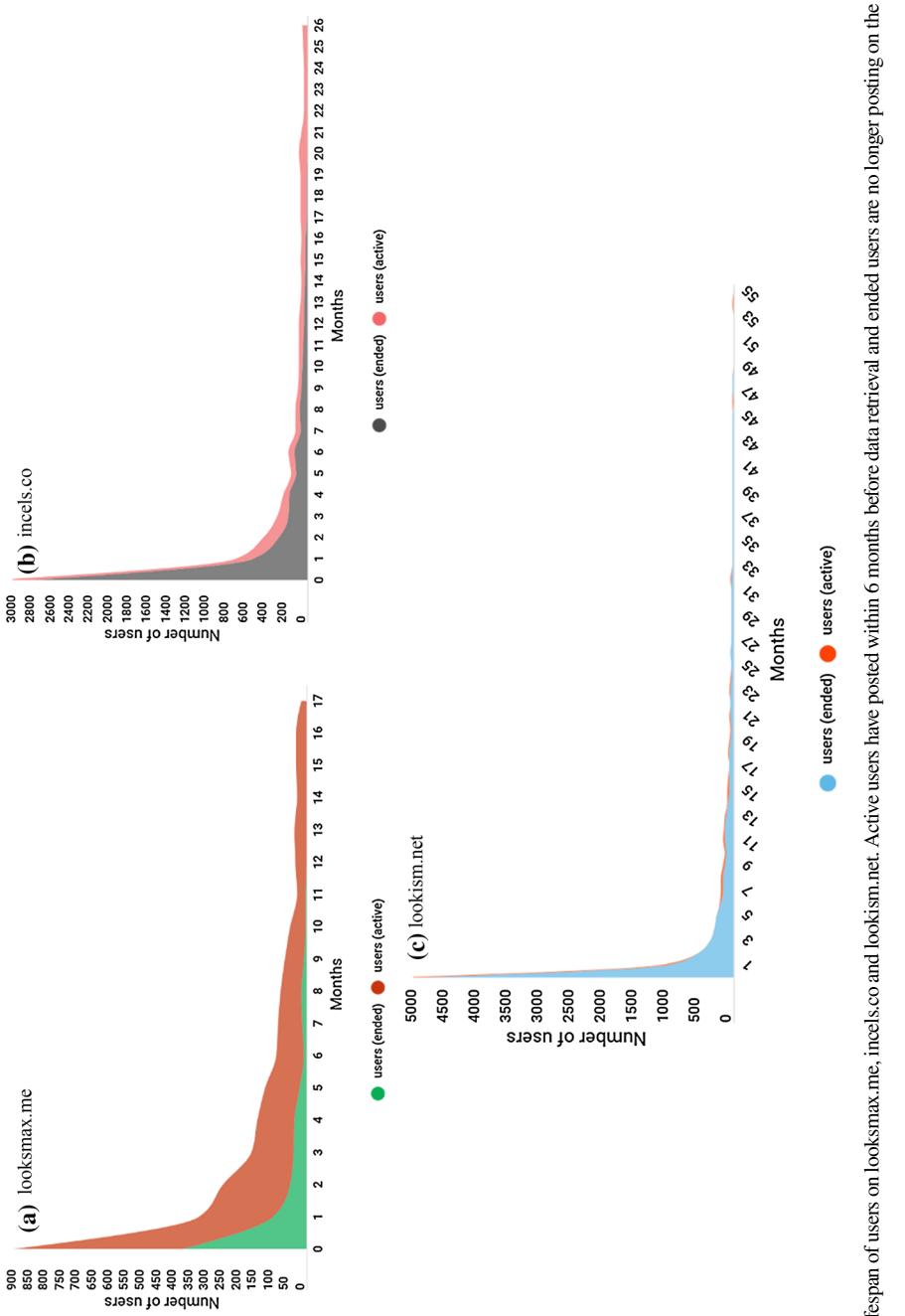

Fig. 3 Lifespan of users on looksmax.me, incels.co and lookism.net. Active users have posted within 6 months before data retrieval and ended users are no longer posting on the forums

users outnumber inactive users in all cohorts. Over 500 users had started posting in the month before our data retrieval, the fastest growth among the forums at the time.

Limitations of the study

We have already described several limitations of this study in Sect. "Methods": The sheer amount of text to be analyzed combined with the inadequacies of even state of the art methods in automated text analysis inevitably add a measure of uncertainty to all results. We have tried to ameliorate this by choosing large representative sample sizes for the automated analysis (see Sect. "Measuring the prevalence and distribution of toxic language") and by using a classifier that has been shown to deliver relatively stable performance across different domains (Berglind et al. 2019).

The lack of strict criteria for judging whether a given text features toxic language provides some leeway for subjectivity. As the classifier has been trained on data (and judgments) from different sources and annotators, we estimate it to represent an average understanding of toxicity. We used this classifier in a uniform fashion across our data to ensure that all toxicity level measurements of the different forums adhere to the same standard.

The manual annotation of the toxicity categories (see Sect. "Measuring the prevalence and distribution of toxic language") is very time-consuming, and the sample sizes had to be smaller.

A more general limitation is that we focus on the analysis of text. We cannot know for certain the motivations of the writers. The members of the incel forums are very much aware of their posts being read by law enforcement, journalists and scientists, and we have encountered cases where members took pride in having been quoted in articles. We cannot rule out that writers sometimes use drastic language and topics merely to shock the audience of outsiders.

Discussion

Incel forums seem to fill a critical need, and meeting others in the same situation that can validate one's feelings may contribute to a sense of belonging. For many who define themselves as incels, it is probably a significant and positive experience to be part of a community, one that understands both the grief and the rage that constant rejection experiences can arouse. The forums also offer explanations as to why some people face constant rejections from women. However, in the long run, the incel forum culture becomes destructive since members fuel one another's depression, rage, and appearance fixation instead of supporting each other to mature and develop. The world outside the forums is painted as hostile and inhospitable, while small details of someone's appearance are portrayed as insurmountable obstacles. The instrumental view of sex and relationships and the negative picture of women increase the risk of making it even more difficult to relate to women.

Several deadly attacks have been conducted by people who see themselves as incels. The perpetrators were looking for revenge, not on one person, but on all

women. We noticed in our analysis that hate towards women was the most common direction of toxic language. This shows that the destructive environment of the incel forums affects significantly more than just the forum users.

Incels are similar to political extremists in their attitudes to the outside world (Baele et al. 2019). There is a distinct mentality of Us-vs-Them, notions of dominance hierarchies, and rigid, dogmatic notions of how the world is organized. But unlike political extremists, incels have no common goal, no utopia to strive for. Some want to return to a more patriarchal society, but many have no vision for society. Adherents of the black pill philosophy believe that evolution and human nature have sentenced them to lifelong suffering, while many other incels mainly focus on their own situation. Another difference is self-contempt. Political extremists often have a grandiose, bloated self-image in which they see themselves as good, insightful, and courageous (Zavala and Lantos 2019). For incels, the reverse usually applies: They tend to hate and despise themselves. In our analysis, we noticed that toxicity towards themselves differed among the three forums. On incels.co it was the second most common form of toxicity (after misogyny), while on looksmax.me and lookism.net toxic language involving ethnicity and other forum users was more common.

Our analysis of lifespans for the different incel forums showed that most users are only active for a short time period on the forums. Our lifespan measurement method is limited in that it cannot provide information about members or visitors who merely read. However, none of the forums require membership registration to read the posts. The users in our lifespan statistics had gone this additional step of registering, which indicates an interest beyond mere curiosity, yet still their vast majority ceased to contribute to the forums after only a short while. This could be interpreted as a promising development, though the exact reasons for the short durations are not known at this point.

While the lifespan can give us information about the time period that individuals are active in the incel environment it does not tell us anything about the level of toxic language among the individuals. For future research it would be interesting to study the development of toxic language of new members. According to Giles et al. (1991), people engaged in conversation subtly and non-consciously tend to shift their speech patterns so that they will become more similar to the speech patterns of the conversation partner. Such linguistic accommodation is partly motivated by a desire to gain others' social approval, and thus used more by those who have a stronger such desire, i.e., persons with lower status, or persons who have not (yet) established their identity and status in a group or dyad they wish to be part of Giles et al. (1991). When an individual starts participating in discussions on a discussion forum it is common that they will adapt to the linguistic style and norms of the forum. Studies of other online forums have found that when new members join, their language use is highly different to that of the more established members but over time their language adapts to the forum until at some point it eventually consolidates (Danescu-Niculescu-Mizil et al. 2013).

One question arises after studying the toxic environment on incel forums: How do we relate to the incel forums' hateful messages and their links to violent acts? It is clear that it does not help to shut down the forums. Incel ideology has mostly been

driven away from large social media, so moderation is effective in general, but the current dedicated incel forums are safe spaces that explicitly allow and encourage such toxic messages. They are maintained by technologically competent individuals who are sympathetic to the incel subculture, and if shut down these forums will just appear somewhere else under a new name. It would be virtually impossible to eliminate the spreading of incel ideology without endangering our open society. We must adapt to a reality where such toxic online environments continue to operate, and we need to focus on ways to reduce the harm they cause. Long-term reduction of harm involves more research to assess the extent to which the incel culture is in itself a threat, what attracts individuals to engage in it, and what causes some of these individuals to turn their violent fantasies into action. Obtaining a better understanding of the sociopsychological mechanisms behind incel activity is necessary to be able to present alternative narratives and encourage deliberation rather than affective polarization. As several perpetrators have expressed overlapping motives, it should also be investigated to what extent there is a link between the incel culture and other violence-promoting environments. Short-term harm reduction includes developing methods for the risk assessment of online content and maintaining a continuous awareness of the ever-changing landscape of incel forums.

Funding Open access funding provided by Swedish Defence Research Agency. This work was funded by the Swedish government.

Data availability The data was retrieved from open internet forums and is thus available to other researchers.

Code availability We do not make our code available, but the framework used to train our classifier is publicly available, as is the data (see above), making it is feasible to replicate our findings.

Declarations

Conflict of interest Not applicable.

References

Anti-Defamation League (ADL). (2018). The Online Hate Index. Retrieved from Anti-Defamation League: https://www.adl.org/resources/reports/the-online-hate-index

Ashcroft, M., Fisher, A., Kaati, L., Omer, E., & Prucha, N. (2015). Detecting Jihadist Messages on Twitter. In J. Brynielsson, & M. H. Yap (Ed.), *2015 European Intelligence and Security Informatics*

Conference, Manchester, United Kingdom, September 7-9, 2015 (pp. 161–164). IEEE Computer Society

Baele SJ, Brace L, Coan TG (2019) From "Incel" to "Saint": Analyzing the violent worldview behind the 2018 Toronto attack. Terrorism and Political Violence. https://doi.org/10.1080/09546553.2019.1638256

Beauchamp, Z. (2019). Our Incel Problem: How a support group for the dateless became one of the internet's most dangerous subcultures. https://www.vox.com/the-highlight/2019/4/16/18287446/incel-definition-reddit

Berglind, T., Pelzer, B., & Kaati, L. (2019). Levels of hate in online environments. In F. Spezzano, W. Chen, & X. Xiao (Ed.), *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019* (pp. 842–847). ACM. https://doi.org/10.1145/3341161.3343521

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017). The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3175–3187). New York, NY, USA: ACM. https://doi.org/10.1145/3025453.3026018

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: user lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web (WWW '13)* (pp. 307-318). New York: ACM

Davidson, T., Warmsley, D., & Macy, M. W. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International AAAI Conference on Web and Social Media*. Montreal, pp. 512–515.

Golec de Zavala A, Lantos D (2019) Collective narcissism and its social consequences: the bad and the ugly. Curr Dir Psychol Sci 29(3):273–278

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Dynel M (2020) Vigilante disparaging humour at r/IncelTears: humour as critique of incel ideology. Lang Commun 74:1–14

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. M. (2018). Hate Lingo: a target-based linguistic analysis of hate speech in social media. *CoRR, abs/1804.04257*. http://arxiv.org/abs/1804.04257

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring Misogyny across the Manosphere in Reddit. *WebSci '19 Proceedings of the 10th ACM Conference on Web Science*, (pp. 87–96). http://oro.open.ac.uk/61128/

Fernquist J, Kaati L, Pelzer B, Cohen K, Akrami N (2020) Hope, cope & rope. FOI, Stockholm

Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv. https://doi.org/10.1145/3232676

Frenda S, Ghanem B, Montes-y-Gómez M, Rosso P (2019) Online hate speech against women: automatic identification of misogyny and sexism on Twitter. J Intell Fuzzy Syst 36:4743–4752

Giles H, Coupland J, Coupland N (1991) Accommodation theory: communication, context, and consequence. In: Giles H, Coupland J, Coupland N (eds) Contexts of accommodation: developments in applied sociolinguistics. Cambridge University Press, New York, pp 1–68

Ging D (2019) Alphas, betas, and incels: theorizing the masculinities of the manosphere. Men Masc 22(4):638–657

Glace AM, Dover TL, Zatkin JG (2021) Taking the black pill: an empirical analysis of the "Incel." Psychol Men Masc 22(2):288–297

Hatebase. (2020). *Hatebase*. https://hatebase.org/

Hoffman B, Ware J, Shapiro E (2020) Assessing the threat of incel violence. Stud Confl Terror 43(7):565–587

Holmberg, C. (2020). Mäns hat mot kvinnor—ett växande terrorhot. *Sveriges Radio*. https://sverigesradio.se/sida/artikel.aspx?programid=83&artikel=7366942

Horta Ribeiro, M., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., & Greenberg, S. Z. (2020). *The Evolution of the Manosphere Across the Web.* arXiv: 2001.07600

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *CoRR, abs/1702.08138*. Hämtat från http://arxiv.org/abs/1702.08138

Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M., & Akrami, N. (2018, 5). Monitoring Targeted Hate in Online Environments. In E. Lefever, B. Desmet, & G. D. Pauw (Ed.), *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018). Paris: European Language Resources Association (ELRA).

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (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.

Jigsaw, & Google. (2020). Perspective API. *Perspective API*. Retrieved from https://www.perspectiveapi.com/

Kaati, L., Omer, E., Prucha, N., & Shrestha, A. (2015). Detecting Multipliers of Jihadism on Twitter. *IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14–17, 2015* (pp. 954–960). IEEE Computer Society.

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (p. 1621â€“1622). Bellevue: AAAI Press.

Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. *CoRR, abs/1712.06427*. http://arxiv.org/abs/1712.06427

Maxwell D, Robinson SR, Williams JR, Keaton C (2020) "A short story of a lonely guy": a qualitative thematic analysis of involuntary celibacy using reddit. Sex Cult 24(6):1852–1874

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (p. 3111â€“3119). Red Hook, NY, USA: Curran Associates Inc.

Moonshot. (2020). Incels: A Guide to Symbols and Terminology. *Incels: A Guide to Symbols and Terminology*.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The Effect of Extremist Violence on Hateful Speech Online. *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25–28, 2018* (pp. 221–230). AAAI Press.

Papadamou, K., Zannettou, S., Blackburn, J., Cristofaro, E. D., Stringhini, G., & Sirivianos, M. (2020). Understanding the Incel Community on YouTube. *CoRR, abs/2001.08293*. https://arxiv.org/abs/2001.08293

Pelzer, B., Kaati, L., & Akrami, N. (2018). Directed Digital Hate. *International Conference on Intelligence and Security Informatics* (pp. 205–210). IEEE.

Preston K, Halpin M, Maguire F (2021) The black pill: new technology and the male supremacy of involuntarily celibate men. Men Masc. https://doi.org/10.1177/1097184X211017954

Pyke KD (2010) What is internalized racial oppression and why don't we study It? Acknowledging racism's hidden injuries. Sociol Perspect 53(4):551–572

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Ed.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, (pp. 6–9).

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2018). Hateminers : Detecting Hate speech against Women. *CoRR, abs/1812.06700*. Retrieved from http://arxiv.org/abs/1812.06700

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, 7). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Florence: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.* Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Texas Fusion Center. (2020). *Texas Domestic Terrorism Threat Assessment.* Intelligence & Counterterrorism Division, Texas Department of Public Safety