Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

From Isolation to Desolation: Investigating Self-Harm Discussions in Incel Communities

Moonis Ali[1], **Savvas Zannettou**[2]
[1]Max Planck Institute for Informatics
[2]Delft University of Technology
moonis.ali@mpi-inf.mpg.de, s.zannettou@tudelft.nl

Abstract

Incel communities have recently attracted the public's interest mainly due to their high degree of extreme views and involvement in real-world violence. A common theme in Incel communities is self-harm discussions. Despite this, beyond small-scale qualitative analyses of self-harm discussions in Incel communities, we lack a large-scale quantitative understanding of how Incels discuss self-harm and how it differs from mainstream communities. In this work, we aim to demystify self-harm discussions in Incel communities using a data-driven approach and understand how Incels differentiate from mainstream communities. We use a dataset of 6.4M posts from 18 Incel subreddits and 2.4M posts from an Incel forum, as well as 5.8M posts from two mainstream subreddits discussing mental health. Using word embedding approaches, temporal analyses, topic modeling, and qualitative analysis, we shed light on self-harm discussions in Incel and mainstream communities and their evolution over time. We find substantial differences in the language related to self-harm deployed among the communities; we find that Incels use niche terms related to self-harm, which is not the case in mainstream communities. We observe that over time, language related to self-harm evolves considerably more among Incels than in mainstream communities. Also, we observe that negative perception of their Physical Appearance is the most recurrent theme in self-harm conversations for Incels, which does not feature in mainstream communities. Finally, by analyzing social factors, we find that Substance Abuse is the most closely associated social factor to self-harm in Incel and mainstream communities and that Physical Appearance, over time, is becoming increasingly closely related to self-harm discussions in Incel communities.

1 Introduction

The Web is a diverse ecosystem comprising a conglomeration of Web communities. Within online communities, certain groups of individuals or cultures are often categorized as "niche" (Zannettou et al. 2020; Ribeiro et al. 2020) in terms of their unique sociocultural characteristics. These mediums afford individuals space for expressions and views that can range from controversial and conspiratorial to deeply personal and intimate discussions. One such example is the *Incels* community. The term "incel" was originally coined to create a community where lonely people of all genders who suffer from problems to foster romantic relationships could support each other (Brunt et al. 2021). However, over time they are now considered an extreme and niche community, wherein Incels (Involuntary Celibates) are mostly young men often categorized to be lonely, misogynistic, and open to committing violence (Ribeiro et al. 2020; Broyd et al. 2022; Daly and Laskovtsov 2021). While Incels represent a diversity of individuals, they often reflect a propensity to inflict harm (to themselves or others) (Brunt et al. 2021).

The Incels community has attracted significant attention due to a troubling pattern of mass violence (Townsend 2022). Many instances of violence, dating back to 2014, have been ascribed to individuals identifying as Incels (Broyd et al. 2022; Lopes 2023; Hoffman, Ware, and Shapiro 2020). A common element in most of these instances of violence is that the perpetrators end their own lives after committing violent acts. While our knowledge of Incels' demographics mainly relies on online polls and surveys conducted by the community itself (Broyd et al. 2022), recent survey results indicate that all respondents were young males aged 18 to 30, living with their parents, and lacking intimate experiences (ADL 2023). The majority expressed discontent with their lives, with 95% subscribing to the blackpill ideology, which represents the most extreme form of Incel identity and is associated with self-harm or harm to others (Brunt et al. 2021). Furthermore, 68% of respondents acknowledged experiencing depression, 74% reported anxiety, and 40% disclosed having been diagnosed with autism. Overall, previous mass violence events and survey results emphasize the problem of self-harm and mental health issues that exist within the Incel community.

Previous work has taken an interest in understanding Incel communities. Most previous data-driven works, however, focused on studying misogyny/toxicity (Farrell et al. 2019; Gothard et al. 2021; Ribeiro et al. 2020; Pelzer et al. 2021), characterizing Incel content (Papadamou et al. 2020) and identifying Incels (Hajarian and Khanbabaloo 2021). Other research has taken an interest in understanding self-harm discussions in Incel communities, with most of the work focusing on qualitative analysis of smaller-scale datasets (Daly and Laskovtsov 2021; Brunt et al. 2021; Blondel et al. 2008). Overall, as a research community, we lack large-scale data-driven analysis of how Incels are discussing self-harm and

methods to demystify these online phenomena.

Research Questions. In this work, we aim to address this research gap and deepen our understanding of Incel self-harm discussions, the problem's nuances, and how discussions change over time. We focus on:

- **RQ1:** What terminology related to self-harm is used by Incel and mainstream communities? Does self-harm language change over time?
- **RQ2:** What specific methods and drugs for self-harm are discussed in Incel and mainstream communities? Do mentions of methods and drugs change over time?
- **RQ3:** What are the different social factors associated with self-harm in Incel and mainstream communities? Does the relationship of different social factors with self-harm vary over time?

To answer our questions and effectively compare discussions related to self-harm among Incels with popular mainstream mental-health communities, we utilize datasets containing 14.6M posts from multiple Incel communities and two mainstream subreddits focusing on discussing mental health. Then, we use methods to discover and analyze discussions related to self-harm and their evolution over time, including Word2vec and visualization techniques to associate words, create lexicons based on the semantic similarity of words, perform topic modeling/thematic analysis, as well as undertake a social factors analysis of self-harm terms.

Main findings. We find the following:

- We find that Incels use niche terms such as "sui" (substitute for suicide) in the context of self-harm, a behavior that does not exist in mainstream communities discussing mental health issues. Also, we find that mainstream communities speak of self-harm in more diverse contexts than Incels, e.g., ideation, whereas, Incels mostly speak of self-harm in violent contexts. From the temporal analysis of self-harm language, we observe that the language related to self-harm among Incel communities changes notably over the years, with niche terms being introduced in the later years. Conversely, the mainstream communities use a more consistent language over the years. (**RQ1**).
- We find substantial differences in the use of self-harm methods and drugs between Incel and mainstream Web communities. We find that Incels use niche terms for methods of self-harm like "ering," "rot," and "rope," which is not the case for mainstream communities. The difference in usage of drugs used for self-harm is even more stark. Mainstream communities mainly mention anti-depressant drugs, while Incel communities do not. Also, we find that Incels mainly focus on discussions to identify "painless" methods for self-harm (e.g., chemicals such as helium or nitrogen), while mainstream communities talk about alcoholism, side effects of anti-depressants, and suicide using guns. Temporal analysis of self-harm methods and drugs revealed that the posting activity for posts mentioning self-harm methods and drugs does not substantially change across the years for any of the mainstream or Incel communities (**RQ2**).
- We find that social factors like Substance Abuse and Financial Troubles are associated more with self-harm for both baseline and Incel communities. Temporal analysis of social factors reveals that the association of Physical Appearance with self-harm undergoes the sharpest increase over the years in Incel communities. (**RQ3**).

Disclaimer: We study self-harm in Incel communities; content included in this paper might be disturbing.

2 Related Work

2.1 Quantitative Perspectives on Incels

Most of the interest received by Incels from computational social science researchers has been related to the study of the "Manosphere." The Manosphere can be termed as a group of online communities bound by a common theme, i.e., issues faced by men (Dewey 2014). Ging (2017) argues that the power of online Web communities to facilitate the spread of information across platforms without any "boundaries" has led to an increase in the spread of extreme views, often characterized by misogyny and violent rhetoric. Incels are often considered extreme and violent groups within the wider Manosphere; motivated by this, Farrell et al. (2019) undertake an analysis of Manosphere subreddits that contain mostly Incel subreddits. By focusing on misogyny, they develop a lexicon from dictionaries that are related to violence, harassment, etc, and then quantify their prevalence on Reddit. Ribeiro et al. (2020) undertake a large-scale analysis of how the Manosphere has evolved between the mid-2000s to 2019, by analyzing data from 51 subreddits and six forums. By observing user participation in various Manosphere communities and the content of their posts, they conclude that communities like Incels are overtaking older communities such as Pickup Artists and Men's Rights Activists. They also show that the newer communities are more toxic and misogynistic when compared to older communities within the Manosphere. Gothard et al. (2021) analyze 3.5M posts from subreddits associated with Incels. Using rank-turbulence divergence to compare word rank distributions of Incels' posts to non-Incel communities' posts, they finalize a lexicon of coded misogyny and violent terms. They propose that these "cryptolects" can help researchers to study niche misogynistic and extreme language phenomena among Incels. Also, Pelzer et al. (2021) try to understand the differences that exist within Incel communities. They study a large dataset of 8.5M posts extracted from three popular online Incel forums. They show that certain Incel forums can be more toxic than others and that Incel forums differ in selecting toxic speech targets.

2.2 Self-harm Discussions in Incel Communities

Most of the work that tries to understand and analyze self-harm-related themes in Incel communities uses manual, survey-based, or qualitative methodologies. Broyd et al. (2022) present a qualitative review of existing psychological and socio-psychological literature featuring Incel communities. They discover that mental health disorders and the propensity to self-harm are recurring features of the Incels. By observing the mental health history of Incel attack perpetrators, they further elucidate the association between mental disorders and violent Incel attacks. They also quote ev-

idence that suggests that people with pre-existing depression and suicidality could exacerbate such conditions after *"becoming a part of Incel communities"*. They also make clinical risk assessment recommendations such as considering "suitability" for the sale of gun licenses and potential weapons such as knives, poison, and acids to Incels. Stijelja and Mishara (2022) present a narrative review of the psychosocial characteristics of Incels. They observe that suicide and the glorification of suicide are highly prevalent themes among Incels. They also state that there exists a high number of users who claim that they have considered suicide. They conclude by calling for more clarity through a deeper analysis of the subject that is based on empirical data. They mention that a better understanding of the psychological and social features of Incels is required before targeted interventions are designed to address specific problems. Moskalenko et al. (2022) ran a large-scale survey on the Incels.co forum between December 2020 and January 2021. In the survey, 311 individuals participated, barring 17, all identified as Incels. They report that almost all Incels reported a history of bullying or persecution. Most reported struggles are associated with mental health issues. They also propose that more research needs to be undertaken to identify Incels with mental health markers so that effective interventions are developed to help them. Daly and Laskovtsov (2021) provide a more fine-grained qualitative study of self-harm among Incels. They manually study 80 suicide posts compiled by u/IncelGraveyard (a Reddit user) across various Incel subreddits. They note that 70% of posts mention the "methods" that the individuals planned to use for their suicide. They further note mentions of specific terms such as "sui" (term for suicide), "rope," which they observe is used as a verb that means to *"kill oneself using a rope"*. They also point out how certain Incels plan to use chemicals like "Potassium Dichromate," "Sodium Thiopental," and others to commit suicide. Further, they also point out mentions of guns among some Incels as a "masculine" way to self-harm.

Remarks. We observe differences between the focus of previous data-driven works and qualitative works that study Incels. While data-driven works focus on themes like misogyny, extremism, and violence, qualitative work examines evidence that relates to self-harm within the Incel community on a small scale. As manual efforts can help to gain some idea to demystify self-harm, there is an obvious caveat: manual efforts simply do not scale. It is not possible for manual efforts to keep up with the rate at which content is posted across platforms such as Reddit, large online forums, etc. This is the gap that this work aims to fill; we attempt to develop reproducible and scalable data-driven techniques to study this understudied phenomenon.

3 Dataset

To understand mental-health-related issues in Incel communities and how they compare with mainstream communities, we leverage and complement datasets made available from previous work. Particularly, for the Incel-related communities, we use data from Ribeiro et al. (2020) that made available a large-scale dataset of posting activity in Incel subreddits and an Incel Forum. The Incel Subreddits dataset includes data from 18 Incel-related subreddits, including 6.4M posts shared between September 2010 and April 2019. The Incel Forum dataset includes data from "Incels.is," a popular forum used for Incel-related discussion, and includes 2.4M posts shared between November 2017 and June 2019. For the mainstream communities, we elect to use two popular and mainstream subreddits that focus on discussions related to mental health issues, particularly r/depression, and r/SuicideWatch. We select these two subreddits as our baselines, as previous work focusing on similar phenomena used these communities (De Choudhury et al. 2016; Alambo et al. 2019). We collect all submissions and comments made to these subreddits from the inception of the subreddits until the end of April 2019, using the monthly dumps made available from Pushshift (Baumgartner et al. 2020). The mainstream datasets include 3.9M and 1.9M posts from r/depression and r/SuicideWatch, respectively. Note that the monthly dumps from Pushshift are no longer publicly available due to changes on Reddit's API and Terms of Service; despite this fact, we still believe that analyzing these large-scale datasets to study phenomena like self-harm discussions is important for Computational Social Science community. To ensure that our results are reproducible, we will make our dataset available to researchers upon request.

4 Methodology

4.1 Semantic Association & Visualization

We train and utilize Word2vec models (Mikolov et al. 2013) to assess the semantic similarity between terms based on their word embeddings. Following Zannettou et al. (2020), we use Word2vec models to find and visualize niche terms related to self-harm. Also, we follow the work by Hamilton, Leskovec, and Jurafsky (2016) to study semantic change over time using Word2vec models trained on a per-year basis. We use Word2vec for assessing word semantics mainly because it is an unsupervised and fast method that encapsulates how words are used in a large corpus of data, hence allowing us to assess how users are discussing self-harm.

Training Word2vec models We train a Word2vec model for each data source; one for all the Incel subreddits, one for the Incel Forum, and one for each of the two mainstream subreddits. We train Word2vec skip-gram models with an embedding dimension of 300 and sliding window (context size) of 7, considering words that occur at least 100 times in the dataset, following the method by (Zannettou et al. 2020). In the subsequent sections, we will refer to trained models using data from all years as *aggregate models*. Our aggregate Word2vec models capture a vocabulary of 23,034 words for Incel Reddit and 19,105 for the Incel Forum. For the mainstream baselines, our aggregate Word2vec models capture a vocabulary of 20,411 words for r/depression and 14,399 for r/SuicideWatch. To understand temporal semantic changes among words related to self-harm across the years, we also train separate Word2vec models for each community and each year present in our datasets. Note that when training temporal Word2vec models, we exclude data before 2014 (i.e., we do not train Word2vec models before 2014) as the

number of posts before 2014 is small (less than 25K monthly posts in each community). We use the same configuration, as mentioned above when training the yearly Word2vec with the only difference, that we consider a word to be part of the vocabulary if it appears at least 50 times; we make this modification given the fact that for the earlier years, we do not have large volumes of data. We refer to the yearly Word2vec models as *temporal models.*

Model Alignment for Temporal Analysis We aim to analyze the evolution of language in Incel and mainstream baselines over multiple years using our temporal Word2vec models. To compare Word2vec models that are trained on different datasets, we must ensure that the vector spaces of different models are comparable, i.e., the resulting vector spaces are not independent with different orientations and/or axes alignment. Thus, direct comparisons between word (vectors) from different spaces would not be meaningful. The vector spaces must be aligned to make meaningful comparisons so that the same word has a comparable position across the different vector spaces given by different temporal models. This alignment process adjusts the vector representations from different models to a comparable space, allowing for comparing semantic changes over time (Hamilton, Leskovec, and Jurafsky 2016). To achieve alignment for each community, we utilize the "orthogonal Procrustes" alignment method, as shown in one of the pioneering works related to the evolution of language (Hamilton, Leskovec, and Jurafsky 2016). We deploy this method as it ensures consistency of axes for the two vector spaces and preservation of approximate directions and magnitudes of the embeddings representing the same words in two different vector spaces (Botarleanu et al. 2022). We align each of our temporal models with respect to the latest model (2019) to get an idea of how language changes with respect to the latest year. Hereon, we refer to temporal models that are aligned to the latest model (2019) for each community as *aligned models*.

Visualization As we aim to assess the use of words around specific self-harm terms, we use graph theory and community detection to visualize their relationships. Following again Zannettou et al. (2020), we create 2-hop ego networks that are graphs where nodes are terms and terms are connected if they have a similarity above a pre-defined threshold based on their word embeddings. We do this by starting from a seed work and adding terms similar to the seed term and their neighbors. Then, we apply community detection to identify the main word communities based on the graph structure. For instance, starting from the term "suicide," we can visualize all the communities of terms related to discussions of self-harm using the Word2vec aggregate model.

4.2 Lexicons for Self-Harm Methods and Drugs

This work analyzes the methods and drugs mentioned in self-harm discussions. By methods, we mean the actual ways for committing self-harm, e.g., by drug overdose, by gun, etc. Self-harm drugs refer to drugs used for self-harm, e.g., harmful chemicals like fentanyl. To achieve this goal, we create two lexicons that include terms related to methods and drugs for self-harm. In a nutshell, we create the lexicons by using seed lexicons from previous work and other sources and leveraging our Word2vec models to expand the lexicons with community-specific terms in a data-driven manner.

Curating the Self-harm *Methods* Seed Lexicon To create a seed lexicon containing self-harm methods, we leverage three sources: one general source and two Incel-specific. First, we use the work by Daly and Laskovtsov (2021) and from the posts that they survey to relate or refer to committing self-harm, we select three terms for our methods lexicon, specifically, "syringe," "jump," and "jumping." Second, we utilize the Incel wiki[1] especially the Glossary page and pages related to suicide/self-harm related phenomena, which includes terminology used by Incels. We do this to identify terms referring to self-harm. We identify 11 terms referring to self-harm, which we confirm by manually annotating a random sample of 20 posts (from either of our 2 Incel datasets) per term, to ensure that they are used in self-harm discussions. This step is crucial as it adds to our lexicon niche Incel terms like "ldar," "sui," "suifuel," etc. Finally, we complement our seed lexicon with a general source, particularly the work by Yazdavar et al. (2017). This source includes self-harm terms used in mainstream Reddit communities and is based on the PHQ-9 (Kroenke, Spitzer, and Williams 2001) mental health questionnaire formulation. We add another 22 terms including "cut," "od," "overdosing," etc. Ultimately, we combine all three sets of terms and create a seed lexicon of self-harm methods that includes 36 terms.

Self-harm *Drugs* Lexicon To create the self-harm drugs lexicon, we again follow three sources, two general and one Incel specific. First, we use existing medical literature that mentions drugs with the potential for self-harm, such as (Mandour 2012; Sinyor et al. 2012; Druda, Gone, and Graudins 2018). We focus on these works because they pertain to the themes of committing self-harm and suicide using drugs and also as they directly provide names of actual and focused terms that we can use in our downstream tasks. Based on these works, we choose 30 terms such as "codeine," and "morphine." Second, we utilize the self-harm drug terms from the US Centers for Disease Control and Prevention (US CDC).[2] We manually choose 11 terms from the CDC sources, such as "opoid," "cocaine," etc. For the Incel communities, we robustify our lexicon by adding niche Incel drug terms drawing from sources such as (Daly and Laskovtsov 2021) and add terms that specify names of certain drugs such as "nitrogen," and "sodium thiopental," drug terms that they found out that Incels deploy in self-harm contexts. Combining all the terms above, we create a seed lexicon of self-harm drugs consisting of 41 terms. We make all lexicons available upon request.

Semantic Expansion of Lexicons Motivated by the fact that our initial lexicons contain a limited set of terms and are not necessarily tailored for our specific Web communities, we aim to expand the lexicons using the trained aggregate

[1] See https://incels.wiki/w/Main_Page

[2] See https://www.cdc.gov/opioids/basics/terms.html and https://www.cdc.gov/opioids/basics/prescribed.html.

Word2vec models. This expansion allows us to account for the subtle differences among the four communities considered in this work and possibly uncover community-specific terms. To expand our two lexicons (self-harm methods and self-harm drugs) semantically, we take the word embeddings of each term in our respective seed lexicons, and to each (*seed term*), we do a summation with the word embedding of the term "suicide." Then, we manually evaluate the top 20 terms (based on cosine similarity) closer to the aggregate embedding and add the relevant terms to the community-specific lexicon. We repeat this process for all terms in our lexicons and for all four aggregate community-specific Word2vec models. This method allows us to discover new terms related to self-harm that are used in each community in a data-driven way. Note that we chose the term "suicide" for two reasons: first because it directly references self-harm and because the results of RQ1 show that it is used in self-harm contexts. After following the above steps to semantically expand the two seed lexicons, i.e., self-harm methods and self-harm drugs, using the Word2vec models for each of our four communities, we finally have eight lexicons for RQ2. These semantically expanded final eight lexicons consist of four community-specific self-harm methods lexicons and four community-specific self-harm drugs lexicons. The lexicons will be made available upon request.

4.3 Topic Modeling and Thematic Analysis

To achieve a fine-grained analysis of self-harm posts, specifically posts including self-harm drugs and methods (RQ2), we perform topic modeling and thematic analysis of the extracted topical clusters following (Ali and Zannettou 2022). Specifically, we use the BERT-based topic modeling proposed by Grootendorst (2020). We use the BERT-based topic modeling approach as it leverages the power of large-pretrained BERT based models and allows us to capture the semantics of discussions in rich and contextual embeddings at the post-level. The topic modeling approach uses Sentence BERT model (Reimers and Gurevych 2019) to encode posts into multi-dimensional vectors, then reduce the vectors using a dimensionality reduction technique (i.e., UMAP (McInnes and Healy 2018)), and perform clustering using a hierarchical clustering algorithm (i.e., HDBSCAN (McInnes, Healy, and Astels 2017)). Then, each cluster is considered a topic; for each topic, we extract how many posts occur in each cluster and representative posts (i.e., posts closer to the cluster's centroid). Having extracted a set of clusters (i.e., topics), we qualitatively analyze the representative posts using thematic coding analysis (Braun and Clarke 2006) belonging to each cluster and assign them meta-labels. As we can observe each of the topical cluster terms that are considered as descriptions of their respective topics Grootendorst (2020) we evaluate all available representative posts for the clusters. The number of these posts is variable and ranges based on inspection due to the stochastic nature of the clustering generation process.

4.4 Social Factor Analysis

To analyze social factors and how they are discussed in self-harm discussions, we build upon the work of Caliskan, Bryson, and Narayanan (2017), who compared the similarity of terms related to gender and social factors like family and career. They analyze the cosine similarity between the four sets of terms (male, female, family, and career) and analyze which set of terms pertaining to gender is closer to which set of terms related to family and career. We extend this approach with a simple modification suggested by Yu and Fliethmann (2021) that suggest using the centroid of a set of terms for comparison purposes, mainly because it results in more robust comparisons.

Creating Term Sets for Self-Harm and Social Factors Since we aim to capture the semantic relationship between *self-harm* terms and terms related to social factors, we first need to create term sets related to self-harm and social factors. For self-harm, we use the lexicon that includes methods for self-harm, as described in Section 4.2. For social factors, we use term sets for six different social factors; four are inspired from previous work (Caliskan, Bryson, and Narayanan 2017; Arseniev-Koehler and Foster 2022), while two are inspired by our topic modeling analysis (RQ2). Specifically, we use the Career and Family terms released by Caliskan, Bryson, and Narayanan (2017), including eight terms for each social factor. Also, we use terms related to Physical Morbidity (referencing physical health problems) and Financial Troubles (referencing financial distress) made available by Arseniev-Koehler and Foster (2022). We manually select relevant terms provided by (Arseniev-Koehler and Foster 2022); for example, for Physical Morbidity, we use terms like "ailment", "ailing", "diseased" and for Financial Troubles, we use terms such as "poverty", "bankrupt", "debt" etc. We use fourteen terms for Physical Morbidity and nine terms for Financial Troubles. For a broader robust comparison, we create two more additional categories related to Physical Appearance and Substance Abuse. We do this since our RQ2 results demonstrate the prevalence of discussions related to Physical Appearance and Substance Abuse. To curate the set of terms related to Physical Appearance, we use general terms related to aesthetics/beauty, such as "ugly," "beauty," and "attraction." To robustify this set with niche Incel terms, we manually assess the terms mentioned in Incels wiki and add terms related to the Physical Appearance like "looksmaxx" and "gymmaxx." Finally, our Physical Appearance term set contains five seed terms. For Substance Abuse, we utilize terms from the results of our topical modeling (RQ2) from Fig. 3 and Fig. 4 and add seven terms prevalent in themes related to abuse of substances including "alcohol", "sober", "alcoholism" etc. Ultimately, we have six social factor seed lexicons (excluding self-harm).

Since we utilize seed lexicons from diverse sources, these terms may vary depending on their temporal usage and according to the usage of these terms in a specific community. Therefore, for robust curation of semantically relevant terms, we use our aggregate Word2vec and temporal models for each community. We expand the seven sets of seed terms by taking, for each term set, the centroid of the term vectors and adding the twenty closest terms that are cosine similar to this centroidal vector using each respective Word2vec model. The final community-specific semantically expanded

term sets from our temporal and aggregate models can be made available to researchers upon request.

5 Results

5.1 Demystifying Expressions of Self-harm (RQ1)

Here, we aim to demystify and understand expressions that are used by Incels in discussions related to self-harm and identify differences from mainstream communities. To investigate this, we attempt to get closer to the usage of the term "suicide" itself and how it is used in conversations by Incels and mainstream communities. We use our Word2vec models to inspect the terms that are utilized in a semantically similar fashion to our target term, in our case, "suicide." We compare the visualizations of the word embedding space of terms cosine similar to the word "suicide" using a 2-hop ego network and community detection. Fig. 1 shows the resulting 2-hop ego networks starting from the word "suicide." Note that due to space limitations, we only show results from two data sources; the Incels forum and r/depression. Due to legibility concerns of the 2-hop ego networks, we follow (Zannettou et al. 2020) to choose terms to include in the figures. We experiment with multiple cosine thresholds in decrements from 0.9 to 0.4 to include terms in the figures. Finally, we select thresholds of 0.55 for r/depression and 0.49 for the Incels Forum. Put differently, terms are added to this 2-hop ego network if they have a cosine similarity of 0.49 or 0.55 with any other term in the network for the Incels Forum and r/depression, respectively.

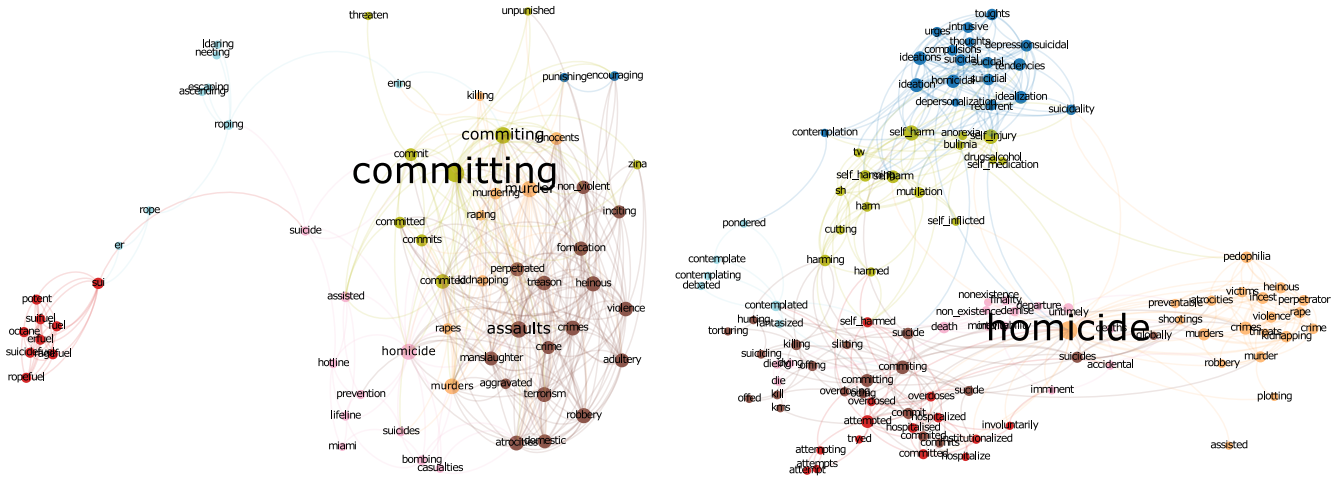
For the Incels forum (Fig. 1a) the red community includes terms that are connected to "suicide" via the niche Incel term "sui" (short form for "suicide"). This community also contains terms such as "potent," "octane," "ragefuel," "ropefuel," "suifuel," "erfuel." Our manual assessment of 20 posts containing each of these terms reveals that half of the terms from this community, specifically "octane," "ropefuel" (fuel/motivation for committing suicide via a rope), "suifuel" (fuel/motivation for suicide) are used by Incels in contexts of suicide and/or suicidal motivation. Similarly, the cyan community containing "er" that refers to Elliot Rodgers (a supposed Incel hero who committed mass violence before committing suicide (Broyd et al. 2022)), "rope," "roping," "ascending," "neeting" and "ldaring" can also be categorized as nuanced Incel expressions based on our manual assessment. The presence of terms such as "er," "ering" (in cyan) and "erfuel" (red) in the figure also raises concerns about the connection of self-harm and mass violence among Incels as indicated by (Brunt et al. 2021), (Joseph et al. 2021) and as described in (EU 2021), the promotion to "go ER" means to commit murder-suicide as a way to inflict punishment on society for an individual's inceldom.

By comparing these results with the mainstream communities (Fig. 1b), we notice a different use of self-harm terms. We find word communities with terms for ideation of self-harm, like the light blue community, including "pondering," "contemplate," "fantasized" and the dark blue community, including "ideation," "suicidality," "thoughts," "urges" etc. We note that this behavior is not the case for Incels. Also, we do not find any niche word communities associated with the term "suicide" like the ones observed in Incels.

Take-away. We find that Incels use niche terms in discussions related to self-harm that do not exist in mainstream communities. At the same time, we find that mainstream communities use a lot of terms related to the ideation of self-harm that do not appear in Incel communities.

Temporal Analysis of Self-harm Language (RQ1) Here, we endeavor to observe differences in the language related to self-harm across different years. We use the methods described in Section 4.1 to align temporal models from each community to the model trained for the latest year, 2019. We do this because we aim to understand how language related to self-harm in the communities changed from the earlier years in comparison to the last year. In Fig. 2, we show how the word "suicide" evolves from the earlier years of each community to the last year (2019). For each community, we make a pair-wise comparison of the word vector for "suicide" using *aligned models*. These models are aligned with the temporal model of 2019. From Fig. 2, we can view the cosine similarity score between the "suicide" vector from the *aligned models* (for each year) and the "suicide" vector (from the 2019 model). We add word clouds to the figure, and try to show the top 20 nearest neighbors to "suicide." The size of each neighboring term is based on its similarity to "suicide." For better readabilty, we skip some terms based on lower similarity scores but we show all terms that that do not feature in previous years (highlighted in red), we do this, to visualize semantic change of terms directly associated with "suicide" over the years. The word clouds in the figure are made from the temporal models, that are *non-aligned*; this is done because aligned models have an intersecting vocabulary, and our aim is to visualize the neighboring terms (to "suicide") for each year irrespective if they are found in the last year (2019), the model to which all other temporal models are aligned to. From Fig. 2, we observe the notable difference between the evolution of the term "suicide" between the two Incel communities and the two baselines. We can observe that the cosine similarity of the vector for "suicide" inside both Incel communities in the initial years compared to the score of last year (2019) starts from around 0.6 and then steadily increases. However, we can see that, for both baselines r/depression and r/SuicideWatch, the range of similarity scores for "suicide" stays within 0.85 to 0.9 over the years. This shows, language related to self-harm implied by discussions that contained the term "suicide" were comparatively more divergent among Incels than for the mainstream communities. From the word clouds, while we observe that word clouds for all communities contain language related to violence such as 'murder', 'killing' etc., we also observe that, among both Incel communities, more nuanced terms such as "er," "rope," "roping" start to appear as close neighbors to "suicide" only in later years; 2016 for Incel Reddit and 2018 for Incel Forum. For the baselines, we do not see the appearance of any niche terms that are close to "suicide."

Take-away. Compared to baselines, we find that language in Incel communities in earlier years differs considerably from the latest year. In the baselines, on the other hand, we



(a) Visualization of terms similar to "suicide" in Incels Forum.

(b) Visualization of terms similar to "suicide" on r/depression.

Figure 1: 2-hop ego network around "suicide." The size of a node is proportional to its degree, the node's color indicates community membership, and the graph is visualized using a layout algorithm where proximity corresponds to similarity.

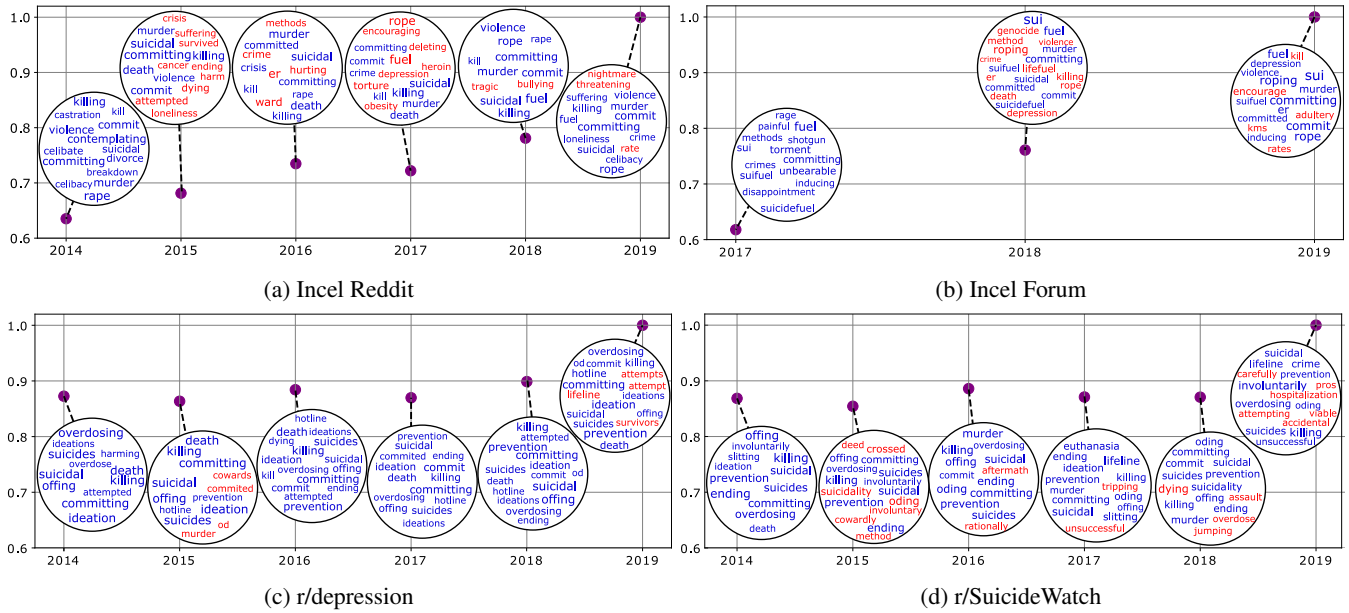
notice less difference in the similarity of how self-harm terminology was used in 2014 than in 2019. Overall, we note that the evolution of self-harm terms among Incel communities is more prominent than in the baselines.

5.2 Understanding Methods of Committing Self-harm (RQ2)

Here, we analyze the methods and drugs for self-harm in discussions in Incels and mainstream communities.

Methods of Self-harm Using the methodology described in Section 4.2, we curate lexicons related to methods of self-harm for each community. For the Incel Reddit dataset, we have curated 161 terms that cover 8.18% of posts, while for the Incel Forum, we have 156 terms with coverage of 15.15% of posts. For the mainstream communities, we have 164 terms for r/depression, which cover 18.39% of posts, and 214 terms for r/SuicideWatch, which feature in 26.87% of posts. We start our analysis by looking into the most popular terms in each community. Table 1 reports the ten most popular terms from our self-harm lexicons. For each term, we report the percentage of posts that include each term over all posts that include any term from our methods self-harm lexicons. For example, in the Incels forums, 74% of posts that include methods of self-harm (based on our lexicons) include the term "er." From this Table, we notice that terms that can be considered niche Incel terms, such as "er," and "ering" feature in the top ten terms in Incels Forum but are missing in Incels Reddit. Similar for the results of RQ1, the occurrence of terms "er" and "ering" indicates not only the semantic closeness of these terms to "suicide" but possibly a connection that closely ties committing suicide and committing acts of mass violence (Brunt et al. 2021). We also note that the expression "rope" is used in both Incel communities and not found in the top 10 most popular terms in any of the mainstream baselines. We can also note that there exists a commonality between the four communities in terms of methods of self-harm with terms such as "off" (a term used for suicide), "cut," "kill," "die," "od" (short form for overdose) occurring in all of them.

To gain a deeper understanding of how methods of self-harm are discussed, we follow the Topic Modeling and Thematic Analysis approach presented in Section 4.3. We apply our methodology to all posts that include any terms from our self-harm methods lexicons for each community. We summarize our results in Fig. 3, where we show the five most popular topics, based on the number of posts, for each community. Each row corresponds to a topic and includes a set of terms that describe the topic (separated by the | symbol), as well as a representative document. To offer visual cues, we use color to indicate similar frequently occurring themes within and across communities (we only color themes if they occur more than once within or across communities). From these illustrations, we can gather insights about the similarities and differences between and within communities. Specifically, from Fig. 3, we observe that the most popular topic in Incel Reddit is related to Self-loathing and (perception of) Negative Physical appearance (NegPa). This finding is consistent with many previous qualitative reports on Incels (Daly and Laskovtsov 2021) and (EU 2021). We see this topic featuring again at the fifth position, where the topic mostly focuses on envy towards tall people and self-loathing about being "manlets," a derogatory Incel slang for short people. The second most popular topic features discussions on Physical Appearance (PA). However, the focus is not self-loathing. Other than these, we notice Hobbies and Parental-loathing as the third and fourth most popular topics. While the Hobbies topic features mostly non-emotive speech, we see more negative and emotive speech in the Parental-loathing topic. The existence of the Hobbies topic is likely due to the use of lexicon terms in other contexts, like in discussions related to online gaming.



(a) Incel Reddit (b) Incel Forum

(c) r/depression (d) r/SuicideWatch

Figure 2: Temporal analysis of self-harm. The Y-axis shows the cosine similarity of "suicide" from each aligned model to "suicide" from the latest (2019) model. Red color means the word did not feature in previous years.

Incels Reddit		Incels Forum		r/depression		r/SuicideWatch	
Term	**%Posts**	**Term**	**%Posts**	**Term**	**%Posts**	**Term**	**%Posts**
od	36.45	er	74.00	od	36.33	od	22.76
tic	18.00	od	8.25	off	10.21	sui	9.33
off	9.58	off	1.69	sui	6.17	kill	6.59
die	4.00	kill	1.24	die	5.21	off	6.36
kill	3.68	die	1.23	kill	5.12	die	5.55
cut	2.93	rot	1.13	hurt	4.09	hurt	3.73
sui	2.49	rope	1.00	scar	3.23	line	3.29
rope	2.09	sui	0.75	arm	2.62	suicidal	3.12
hurt	1.73	cut	0.64	cut	2.43	tie	2.45
rape	1.54	ering	0.62	suicidal	2.29	scar	2.36

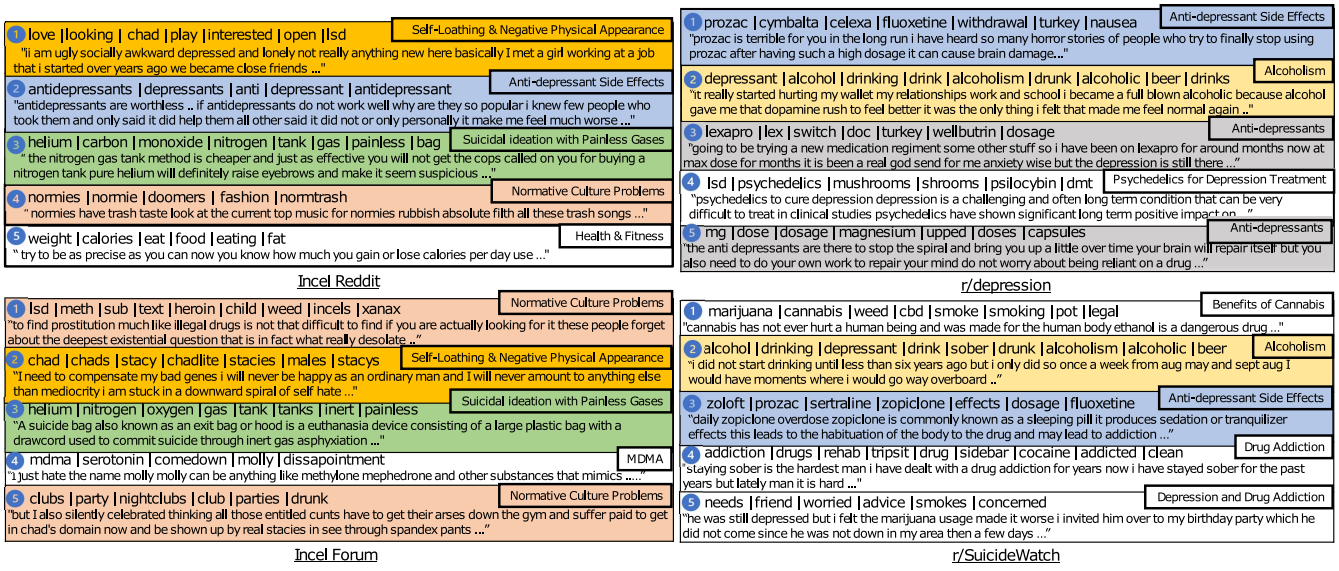
Table 1: Top self-harm methods terms.

For the Incels Forum, the most popular topic is about Suicidal Ideation and NegPA. It features perceptions about NegPA and how lack of interest from women as romantic partners renders young Incels to ideate or/and commit suicide. The second most popular topic features Health and Fitness as the main theme. It talks about working out and eating healthy to attain a better physique. The third topic is about gaming, although not relevant to self-harm. We believe the presence of such a cluster is owed to including terms such as "kill," "dead," etc. in the used lexicons, which are often used for gaming discussions. The fourth most popular topic is about Prostitution as a means to indulge in sexual pleasure for Incels, a phenomenon that in Incel culture is known as "escortcelling." Similar to the results for Incel Reddit, the fifth topic features the theme of Self-loathing and NegPA.

For the baselines (Fig. 3), we can observe that the first two prominent topics in r/depression talk about Depression. The first topic is related to depression in the context of relationship building and romance, whereas the second topic talks about depression in the context of education and the pressure of exams and getting good grades. The third topic features the theme of alcoholism and relates to drinking problems and alcohol addiction and its social cost. The fourth topic is Suicidal Ideation with Guns as a way to commit suicide and the fifth is related to Health and Fitness. For r/SuicideWatch we can notice the prominent occurrence of themes related to Suicidal Ideation, featuring in the three most prominent topics. The first topic concerns Suicidal Ideation in the context of romantic failures and social relationships. The second topic, similar to r/depression, is about Suicidal Ideation with Guns, while the third topic is related to pets and the connection to suicide. The fourth topic, similar to r/depression, is about Alcoholism and how people seek remedies to fight this addiction. Finally, we note the fifth topic, which talks about depression in the context of gender dysphoria and the difficulties that people encounter after the gender transition.

(a) Incels (b) baseline

Figure 3: Topic Modeling and Thematic Analysis of posts including self-harm methods. We color themes that appear more than once in the same community or across different communities.



(a) Incels (b) baseline

Figure 4: Topic Modeling and Thematic Analysis of posts including self-harm drugs. We color themes that appear more than once in the same community or across different communities.

Take-away. We notice a stark difference in the themes discussed between Incels and mainstream communities. Both Incel communities showcase themes related to Self-loathing owed to Negative Physical Appearance and Physical Appearance in multiple contexts. These findings align with previous qualitative works that study Incel's self-harm discussions (Brunt et al. 2021; Daly and Laskovtsov 2021). However, we note this behavior is missing in the baselines, where the focus seems to be more on Alcoholism, Depression in varied contexts, and Suicidal Ideation with Guns. Thus, we see Incels talk more about Physical Appearance, often in negative contexts than the mainstream communities.

Self-harm Drugs Here, we focus on understanding discussions and terms related to self-harm and the use of drugs. For terms related to self-harm drugs, our curated lexicon for r/depression has 134 terms covering 4.22% of posts, and for r/SuicideWatch we have 114 terms covering 3.23% of posts.

Incels Reddit		Incels Forum		r/depression		r/SuicideWatch	
Term	**%Posts**	**Term**	**%Posts**	**Term**	**%Posts**	**Term**	**%Posts**
drug	26.95	meth	62.68	meth	58.04	meth	59.76
drugs	16.12	drug	13.31	pot	5.56	drug	7.70
alcohol	14.51	drugs	8.07	drugs	5.03	pot	6.18
weed	9.69	alcohol	5.72	mg	4.42	drugs	5.02
oxy	3.42	heroin	2.46	alcohol	4.05	alcohol	4.08
depressant	3.15	depressant	0.90	weed	2.96	mg	2.21
heroin	2.79	lsd	0.78	wellbutrin	1.57	weed	1.97
thc	2.75	cocaine	0.58	zoloft	1.55	grams	0.75
depressants	2.51	mdma	0.58	prozac	1.44	heroin	0.69
coke	1.78	xanax	0.47	lexapro	1.10	xanax	0.60

Table 2: Top self-harm drugs terms.

For Incel Reddit, we have curated 76 terms that cover 0.74% of posts, and for Incel Forum we have 67 terms with a coverage of 1% of posts. In Table 2, we report the top ten most used terms from the final self-harm drugs lexicons for each community. We observe that the most popular drug mentioned in all communities, except Incels Reddit, is "meth" with over 58% of all posts that contain any self-harm drug term from our lexicons.

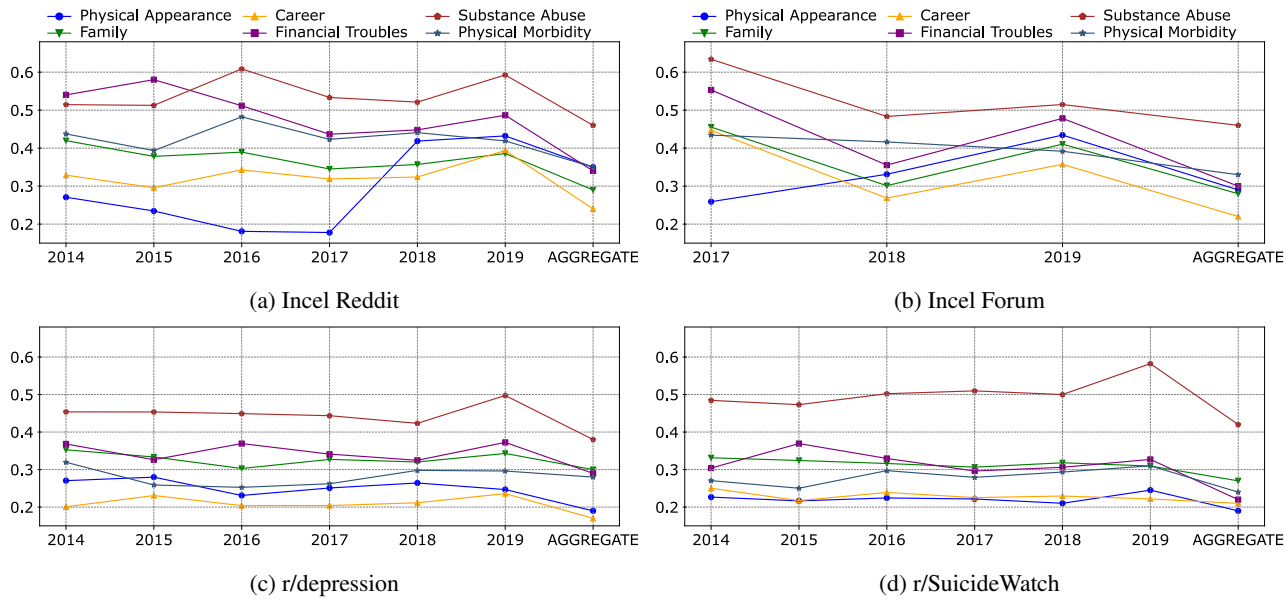
To better understand the use of drug words in self-harm discussions, we perform topic modeling and thematic analysis on the posts containing words from our self-harm drugs lexicons. Fig. 4 summarizes the results of our topic modeling analysis. We observe that for Incel Reddit, the most popular topic relates to Self-loathing due to Negative Physical Appearance (NegPA). The second most popular topic is Side-effects of Anti-depressants and how anti-depressants are bad for health. The third most popular topic features Suicidal ideation with Painless Gases, focusing on committing suicide with carbon monoxide, nitrogen, and helium-like gases as they are perceived as a painless way to commit self-harm. The fourth popular topic, Normative Culture Problems, references lamentations about normative culture. This topic features cynical criticism of the "perceived and prevalent." The fifth topic relates to Health and Fitness. In results for Incels Forum, we see similarity of themes and topics as for Incel Reddit. For the baselines, we observe that for r/depression the most popular topic relates to harmful Side effects of Anti-depressants and features references to "prozac" and "cymbalta" (both popular anti-depressants). The second most popular topic features Alcoholism and the harmful effects of alcohol addiction. The third most popular topic talks about trying out Anti-depressant medications like "lexapro" to alleviate depression. The fourth popular topic references the use of Psychedelics for Depression Treatment. For r/SuicideWatch, we find similar topics with r/depressions with the exception of topics related to the Benefits of Cannabis in treating depression (first topic) and a topic about how drugs can make depression worse and cost social relationships (fifth topic).

Take-away. We observe differences in the themes discussed between Incel and mainstream communities. We see both Incel communities featuring themes related to Self-loathing owed to Negative Physical Appearance, lamentations about Normative Cultural Problems, and Suicidal ideation using Painless Gases. The prevalence of "painless" gases for self-harm aligns with previous qualitative works that study the usage of drugs and chemicals for suicide among Incel communities (Daly and Laskovtsov 2021). We see that all these themes are missing in mainstream communities where the focus seems to be more on Alcoholism, Side-effects of Anti-depressants, and Anti-depressants.

Temporal Analysis of Self-harm Methods and Drugs Using the semantically expanded self-harm methods and drug lexicons for each community, we investigate whether there are substantial changes in the appearance of these terms over time. We search for posts that contain terms from the community-specific lexicons mentioned above. Our analysis shows that for both baselines and Incel communities, there is no substantial difference in the posting activity for posts containing self-harm methods or drug terms over time (we omit the figures due to space constraints).

5.3 Comparing Social Factors Associated with Self-harm (RQ3)

This section presents our results that analyze the connection of multiple social factors with self-harm. For each of the four communities, we calculate the cosine similarity between the centroid of the self-harm word set with the centroids of each of the six curated social factor word sets. To capture the temporal dynamics of how self-harm terms relate to different social factors over time, we utilize our temporal Word2vec models. We also employ the aggregate Word2vec models to capture the overall relationship. Fig. 5 summarizes both the aggregate results (last point in the plot) and the changes over time; it shows the cosine similarity of social factors' centroid with the self-harm centroid using the aggregate and temporal Word2Vec models. We make several interesting observations. First, by looking into the aggregate results, we observe that for all communities, the Substance Abuse social factor is the one that is closer to self-harm discussions with highest cosine similarity for all communities. Second, we observe that for Incel Reddit, the other social factors that are closer to self-harm discussions after Substance Abuse are Physical Appearance (0.35), Financial Troubles (0.34),



(a) Incel Reddit

(b) Incel Forum

(c) r/depression

(d) r/SuicideWatch

Figure 5: Comparison of social factors with self-harm for each community over time. The y-axis shows the similarity of social factors with self-harm across the years and in aggregate.

and Physical Morbidity (0.35), while for the Incel Forum, the second social factor is Physical Morbidity (0.33). Third, for all communities, we observe that the social factor that is furthest from self-harm discussions is Career, indicating that both Incel and mainstream users do not associate Career aspects when discussing self-harm issues. For baseline communities, the same applies to the Physical Appearance social factor (that has approximately the same cosine similarity to Career); on the other hand, this does not apply for Incel communities as the Physical Appearance factor has a substantially higher cosine similarity compared to Career (0.35 vs. 0.24 for Incel Reddit and 0.29 vs. 0.22 for Incel Forum). When looking at the temporal results, we make another interesting observation with regard to Physical Appearance. We observe that in both Incel communities, over time, Physical Appearance is becoming increasingly closer to self-harm discussions. For instance, in Incel Reddit, between 2014 and 2017, Physical Appearance was the last social factor with a cosine similarity that is consistently below 0.3, and in 2018, we observe a substantial increase; this indicates that in recent years, the Incel community on Reddit associated Physical Appearance a lot more with self-harm discussions. A similar trend applies to Incel Forum; over time, Physical Appearance is becoming more associated with discussions related to self-harm. Finally, for the rest of the social factors, we do not observe any substantial shifts in how these social factors are associated with self-harm discussions over time.

Take-away. Substance Abuse is the social factor that is most similar to self-harm over all the years and in aggregate among all communities. Among Incel communities, Physical Appearance shows the sharpest increase over the years in Incel Reddit and remains overall second closest to self-harm. We see Physical Morbidity and Financial Troubles also closely following Substance Abuse. For both baselines, all social factors except for Substance Abuse show less variation in similarity scores between 2014 and 2019.

6 Discussion

Our work and findings yield some important implications for various interested stakeholders, including the research community, operators of social media platforms like Reddit, and professional mental health experts. For the research community, we believe that our work demonstrates how a data-driven approach can provide unique insights into an important problem (self-harm) and the nuances that exist when analyzing niche Web communities like Incels. We demonstrate how interested researchers can apply data-driven methods to analyze communities that are using coded language or niche terms to discuss specific topics (in this case, self-harm discussions). The presented methods are generalizable, and we believe that they can be applied to understanding other online phenomena, as well as other online communities beyond Incels. Also, our research has important implications for social media operators for moderation purposes. Our work highlights that Incels use niche terms for discussing self-harm discussions, which can potentially not be effectively captured by existing moderation approaches. Not only does our work raise awareness about this phenomenon, but we also make available various lexicons pertaining to self-harm terms, which can help existing content moderation procedures better identify self-harm content. Furthermore, we believe that the identification of such self-harm discussions is an important step, which can subsequently be used to train or fine-tune large-language models with the goal of identifying coded or niche terms that are used in self-harm discussions (e.g., see (Mendelsohn et al. 2023)). Finally, we be-

lieve our work, particularly the social factor analysis, can assist professional mental health experts in several ways. First, our analysis can help mental health experts to better understand the group dynamics, peer influences, and social norms within Incel communities, which can assist in more culturally and contextually sensitive intervention strategies. Second, our research demonstrates the social factors and influences that shape discussions related to self-harm; these results can help experts in engaging with these influences to promote positive change within the Incel community. Also, our results can assist experts in tailoring their communication strategies when dealing with Incels, including using language/concepts that resonate with the community and addressing the social pressures and challenges they face.

Limitations. Our work has some limitations. First, we do not capture the full discourse of self-harm discussions because we do not perform any user-level study. Also, our claims are strictly limited to the actual terms associated with self-harm we have examined. At the same time, as with all lexicon approaches, some of the discussions analyzed in this work are false positives and not related to self-harm (e.g., gaming topics might be due to Incels discussing online gaming and using terms like "die" or "kill"). Additionally, the research team does not include a professional mental health expert, and a substantial part of our work relies on manual annotations (e.g., for creating lexicons or selecting relevant terms) that are not conducted by professional mental health experts. All annotations for our work were done by the first author of this work. While the first author is not a professional mental health expert, he has experience in the domain of computational psychology and has worked on similar topics for the past 1.5 years. Despite these limitations, we believe that our work still provides meaningful insights into self-harm discussions on Incel communities.

7 Conclusion

In this work, we performed a large-scale analysis of discussions of self-harm in Incel communities and compared them with mainstream communities on Reddit that focus on mental health. Using data-driven methods based on Word2vec models, we presented methods to identify and visualize words related to self-harm, analyze the topics of discussions in Incels and mainstream communities, and demystify which social factors are closer to discussions of self-harm online. Among other things, our study highlights the substantial differences in how self-harm is discussed in Incel and mainstream Web communities. Also, our work demonstrates how self-harm discussions evolved over time in both Incel and mainstream communities online, highlighting the dynamic landscape of online communities when discussing mental health issues. To assist researchers in further understanding these phenomena, we plan to make available (upon request) resources and lexicons created during our work.

References

ADL. 2023. Online Poll Results Provide New Insights into Incel Community — adl.org. https://www.adl.org/blog/online-poll-results-provide-new-insights-into-Incel-community. Accessed: 2023-05-06.

Alambo, A.; Gaur, M.; Lokala, U.; Kursuncu, U.; Thirunarayan, K.; Gyrard, A.; Sheth, A.; Welton, R. S.; and Pathak, J. 2019. Question answering for suicide risk assessment using reddit. In *ICSC*.

Ali, M.; and Zannettou, S. 2022. Analyzing Antisemitism and Islamophobia using a Lexicon-based Approach. In *ICWSM Workshops*.

Arseniev-Koehler, A.; and Foster, J. G. 2022. Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What it Means to be Fat. *Sociological Methods & Research*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *ICWSM*, 830–839.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008.

Botarleanu, R.-M.; Dascalu, M.; Watanabe, M.; Crossley, S. A.; and McNamara, D. S. 2022. Age of Exposure 2.0: Estimating word complexity using iterative models of word embeddings. *Behavior Research Methods*, 54: 3015 – 3042.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Broyd, J.; Boniface, L.; Parsons, D.; Murphy, D.; and Hafferty, J. D. 2022. Incels, violence and mental disorder: a narrative review with recommendations for best practice in risk assessment and clinical intervention. *BJPsych Advances*.

Brunt, B. J. V.; Brunt, B. S. V.; Taylor, C.; Morgan, N.; and Solomon, J. 2021. The Rise of the Incel Mission-Oriented Attacker. *Violence and Gender*, 8(4): 163–174.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Daly, S. E.; and Laskovtsov, A. 2021. "Goodbye, My Friendcels": An Analysis of Incel Suicide Posts. *Journal of Qualitative Criminal Justice & Criminology*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*, 2098–2110.

Dewey, C. 2014. Inside the 'manosphere' that inspired Santa Barbara shooter Elliot Rodger. https://www.washingtonpost.com/news/the-intersect/wp/2014/05/27/inside-the-manosphere-that-inspired-santa-barbara-shooter-elliot-rodger/. Accessed: 2023-04-23.

Druda, D. F.; Gone, S.; and Graudins, A. 2018. Deliberate Self-poisoning with a Lethal Dose of Pentobarbital with Confirmatory Serum Drug Concentrations: Survival After Cardiac Arrest with Supportive Care. *Journal of Medical Toxicology*, 15(1): 45–48.

EU. 2021. Incels: A First Scan of the Phenomenon (in the EU) and its Relevance and Challenges for P/CVE. https://home-affairs.ec.europa.eu/system/files/2021-10/ran_incels_first_scan_of_phenomen_and_relevance_challenges_for_p-cve_202110_en.pdf. Accessed: 2023-02-06.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *WebSci*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Ging, D. 2017. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, 22(4): 638–657.

Gothard, K.; Dewhurst, D. R.; Minot, J. R.; Adams, J. L.; Danforth, C. M.; and Dodds, P. S. 2021. The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community. arXiv:2105.12006.

Grootendorst, M. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

Hajarian, M.; and Khanbabaloo, Z. 2021. Toward Stopping Incel Rebellion: Detecting Incels in Social Media Using Sentiment Analysis. In *ICWR*.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proc. Assoc. Comput. Ling. (ACL)*.

Hoffman, B.; Ware, J.; and Shapiro, E. 2020. Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism*, 43(7): 565–587.

Joseph, S. M.; Citraro, S.; Morini, V.; Rossetti, G.; and Stella, M. 2021. Cognitive network science quantifies feelings expressed in suicide letters and Reddit mental health communities. *ArXiv*, abs/2110.15269.

Kroenke, K.; Spitzer, R. L.; and Williams, J. B. W. 2001. The PHQ-9. *Journal of General Internal Medicine*, 16(9): 606–613.

Lopes, F. M. 2023. What Do Incels Want? Explaining Incel Violence Using Beauvoirian Otherness. *Hypatia*, 1–23.

Mandour, R. 2012. Antidepressants medications and the relative risk of suicide attempt. *Toxicology International*, 19(1): 42.

McInnes, L.; and Healy, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, abs/1802.03426.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2: 205.

Mendelsohn, J.; Bras, R. L.; Choi, Y.; and Sap, M. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. *arXiv preprint arXiv:2305.17174*.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Moskalenko, S.; González, J. F.-G.; Kates, N.; and Morton, J. 2022. Incel Ideology, Radicalization and Mental Health. *The Journal of Intelligence, Conflict, and Warfare*.

Papadamou, K.; Zannettou, S.; Blackburn, J.; Cristofaro, E. D.; Stringhini, G.; and Sirivianos, M. 2020. "How over is it?" Understanding the Incel Community on YouTube. *CSCW*, 5: 1 – 25.

Pelzer, B.; Kaati, L.; Cohen, K.; and Fernquist, J. 2021. Toxic language in online incel communities. *SN Social Sciences*, 1.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; Cristofaro, E. D.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2020. The Evolution of the Manosphere across the Web. In *ICWSM*.

Sinyor, M.; Howlett, A.; Cheung, A. H.; and Schaffer, A. 2012. Substances Used in Completed Suicide by Overdose in Toronto: An Observational Study of Coroner's Data. *The Canadian Journal of Psychiatry*, 57(3): 184–191.

Stijelja, S.; and Mishara, B. L. 2022. Psychosocial Characteristics of Involuntary Celibates (Incels): A Review of Empirical Research and Assessment of the Potential Implications of Research on Adult Virginity and Late Sexual Onset. *Sexuality & Culture*.

Townsend, M. 2022. Experts fear rising global 'incel' culture could provoke terrorism. https://www.theguardian.com/society/2022/oct/30/global-incel-culture-terrorism-misogyny-violent-action-forums. Accessed: 2023-01-03.

Yazdavar, A. H.; Al-Olimat, H. S.; Ebrahimi, M.; Bajaj, G.; Banerjee, T.; Thirunarayan, K.; Pathak, J.; and Sheth, A. 2017. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. In *ASONAM*, 1191–1198.

Yu, Q.; and Fliethmann, A. 2021. Frame detection in German political discourses : How far can we go without large-scale manual corpus annotation? In *CPSS*, 13–24.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *ICWSM*.

Ethical Statement

Ethics and Broader Perspective. We highlight that our work relies entirely on passively analyzing publicly available data that was made available by previous work. Also, given that the paper focuses on a sensitive topic, we elected to perform all necessary annotations of the data ourselves without using crowdsourcing workers. This minimizes the potential harm and exposure to sensitive content to people outside our research team. When analyzing and presenting the results, we presented aggregate results and ensured that we did not present information and data that can de-anonymize users based on what they posted on Reddit or Incel forums. Also, we did not attempt to track users across websites. Our work demonstrates methods that build on top of previous research work to understand and analyze self-harm discussions in Incel communities. We believe that our work has the merits of demonstrating how we can use data-driven methods to understand and analyze communities that use a lot of coded or niche terminology. Also, our work makes available lexicons that are used by Incels in self-harm discussions upon request. We make them available only upon request to avoid potential harm that may arise by exposing users to niche and potentially harmful terms that are related to self-harm.

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see Section 0 and Section 1
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section 4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No because we use the entire activity on Reddit and the Incel forum for our work.
   (e) Did you describe the limitations of your work? Yes, see Section 7.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethics and Broader Perspective above.
   (g) Did you discuss any potential misuse of your work? Yes, see Ethics and Broader Perspective above.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethics and Broader Perspective above.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...
   NA
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   NA
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see Section 3 and Section 4.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see Section 4.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   NA
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA