frontiers | Frontiers in Social Psychology

TYPE Original Research
PUBLISHED 03 July 2024
DOI 10.3389/frsps.2024.1383152

Check for updates

OPEN ACCESS

EDITED BY
Clemens Stachl,
University of St. Gallen, Switzerland

REVIEWED BY
Rudolf Debelak,
University of Zurich, Switzerland
Jan Digutsch,
University of St. Gallen, Switzerland
Suhaib Abdurahman,
University of Southern California, United States

*CORRESPONDENCE
Miriam Schirmer
✉ miriam.schirmer@tum.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 06 February 2024
ACCEPTED 29 May 2024
PUBLISHED 03 July 2024

CITATION
Matter D, Schirmer M, Grinberg N and Pfeffer J (2024) Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration. *Front. Soc. Psychol.* 2:1383152. doi: 10.3389/frsps.2024.1383152

Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration

Daniel Matter[1†], Miriam Schirmer[1*†], Nir Grinberg[2] and Jürgen Pfeffer[1]

[1]Department of Governance, School of Social Sciences and Technology, Technical University of Munich, Munich, Germany, [2]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beersheba, Israel

This study investigates the prevalence of violent language on *incels.is*. It evaluates GPT models (GPT-3.5 and GPT-4) for content analysis in social sciences, focusing on the impact of varying prompts and batch sizes on coding quality for the detection of violent speech. We scraped over 6.9*M* posts from *incels.is* and categorized a random sample into non-violent, explicitly violent, and implicitly violent content. Two human coders annotated 3,028 posts, which we used to tune and evaluate GPT-3.5 and GPT-4 models across different prompts and batch sizes regarding coding reliability. The best-performing GPT-4 model annotated an additional 45,611 posts for further analysis. We find that 21.91% of the posts on the forum contain some form of violent language. Within the overall forum, 18.12% of posts include explicit violence, while 3.79% feature implicit violence. Our results show a significant rise in violent speech on *incels.is*, both at the community and individual level. This trend is particularly pronounced among users with an active posting behavior that lasts for several hours up to one month. While the use of targeted violent language decreases, general violent language increases. Additionally, mentions of self-harm decline, especially for users who have been active on the site for over 2.5 years. We find substantial agreement between both human coders ($\kappa = 0.65$), while the best GPT-4 model yields good agreement with both human coders ($\kappa = 0.54$ for Human A and $\kappa = 0.62$ for Human B). Overall, this research offers effective ways to pinpoint violent language on a large scale, helping with content moderation and facilitating further research into causal mechanisms and potential mitigations of violent expression and online radicalization in communities like *incels.is*.

KEYWORDS

GPT-4, violent language, Incels, annotator agreement, time analysis

1 Introduction

The term “Incels” (“Involuntary Celibates”) refers to heterosexual men who, despite yearning for sexual and intimate relationships, find themselves unable to engage in such interactions. The online community of Incels has been subject to increasing attention from both media and academic research, mainly due to its connections to real-world violence (Hoffman et al., 2020). Scrutiny intensified after over 50 deaths have been linked to Incel-related incidents since 2014 (Lindsay, 2022). The rising trend of Incel-related violence underscores societal risks posed by the views propagated within the community, especially those regarding women. In response, various strategic and administrative measures have been implemented. Notably, the social media platform Reddit officially banned the largest Incel subreddit *r/incel* for inciting violence against women (Hauser, 2017). The Center for Research and Evidence on Security Threats has emphasized the community's violent, misogynistic tendencies, classifying its ideology as extremist (Brace, 2021). Similarly, the Texas Department of Public Safety has labeled Incels as an "emerging domestic terrorism threat" (Texas Department of Public Safety, 2020).

Incels mainly congregate on online platforms. Within these forums, discussions frequently revolve around their feelings of inferiority compared to male individuals known as "Chads," who are portrayed as highly attractive and socially successful men who seemingly effortlessly attract romantic partners. Consequently, these forums often serve as outlets for expressing frustration and resentment, usually related to physical attractiveness, societal norms, and women's perceived preferences in partner selection. These discussions serve as an outlet for toxic ideologies and can reinforce patterns of blame and victimization that potentially contribute to a volatile atmosphere (Hoffman et al., 2020; O'Malley et al., 2022).

As public attention on Incels has grown, researchers have also begun to study the community more comprehensively, focusing on abusive language within Incel online communities (Farrell et al., 2019; Jaki et al., 2019), Incels as a political movement (O'Donnell and Shor, 2022), or mental health aspects of Incel community members (Broyd et al., 2023). Despite the widespread public perception that links Incels predominantly with violence, several studies found that topics discussed in Incel online communities cover a broad range of subjects that are not necessarily violence-related, e.g., discussions on high school and college courses and online gaming (Mountford, 2018). Nevertheless, the prevalence of abusive and discriminatory language in Incel forums remains a significant concern as it perpetuates a hostile environment that can both isolate members further and potentially escalate into real-world actions.

This paper follows up on how violent content is presented and evolves on *incels.is*, the largest Incel forum. We examine the prevalence and changes in violent content, analyzing specific forms of violence in individual posts and their progression over time at the user level. Our study classifies various types of violent content—explicit vs. implicit, and directed vs. undirected—using both manual labeling and Large Language Models (LLMs). We also assess the effectiveness of OpenAI's GPT-3.5 and GPT-4 models in annotating this content, exploring the challenges associated with these models.

While previous studies have explored the dynamics of violence in Incel forums broadly (cf., Farrell et al., 2019 with a focus on misogyny), there exists a significant research gap in understanding the specific forms of violence articulated in individual posts and the progression of such content at the user level (see the following paragraphs for a more detailed literature review). This distinction is critical as it allows us to determine the extent of violent content on the overall forum level and analyze users' trajectories of posting violent content in their posts, offering insights beyond the collective forum atmosphere.

We initially perform manual labeling on a subset of the data to establish a human baseline and ensure precise categorization for our violence typology, e.g., explicit vs. implicit violence; see Section 5.1. We then employ OpenAI's GPT-3.5 and GPT-4 APIs to classify a greater number of posts, enabling a comprehensive annotation of our dataset. We use the human baseline to assess the performance and ensure the accuracy of the categorization process, and discuss different experimental setups and challenges associated with annotating Incel posts. We then examine how the prevalence of violent content within the forum evolves for each category on the individual and forum levels.

2 Violent language in Incel communities

Within computational social science (Lazer et al., 2009), a diverse body of research has explored the multifaceted landscape of incel posts and forums. Natural language processing techniques have been employed to analyze the linguistic characteristics of Incel discourse, uncovering patterns of extreme negativity, misogyny, and self-victimization. Sentiment analysis, for instance, has illuminated the prevalence of hostile sentiments in these online spaces (Jaki et al., 2019; Pelzer et al., 2021), while topic modeling has unveiled recurrent themes and narratives driving discussions (Mountford, 2018; Baele et al., 2021; Jelodar and Frank, 2021). Other studies have focused on broader communities of misogynistic movements, tracking their evolution over time (Ribeiro et al., 2021a). These studies offer invaluable insights into the dynamics of Incel online communication and serve as a valuable foundation for more comprehensive research to fully understand the complexities of these communities.

Due to misogynistic and discriminating attitudes represented in Incel forums, research focusing on violent content constitutes the majority of academic studies related to this community. Pelzer et al. (2021), for instance, conducted an analysis of toxic language across three major Incel forums, employing a fine-tuned BERT model trained on ∼20,000 samples from various hate speech and toxic language datasets. Their research identified seven primary targets of toxicity: women, society, incels, self-hatred, ethnicities, forum users, and others. According to their analysis, expressions of hatred toward women emerged as the most prevalent form of toxic language (see Jaki et al., 2019 for a similar approach). On a broader level, Baele et al. (2021) employed a mix of qualitative and quantitative content analysis to explore the Incel ideology prevalent in an online community linked to recent acts of politically motivated violence. The authors emphasize that this particular community occupies a unique and extreme position within the broader misogynistic movement, featuring elements that not only encourage self-destructive behaviors but also have the potential to incite some members to commit targeted acts of violence against women, romantically successful men, or other societal symbols that represent perceived inequities.

The rise of research on the Incel community has also shifted the spotlight on users within the "Incelverse," driven by both qualitative and computational approaches. Scholars have embarked on demographic analyses, identifying prevalent characteristics, such as social isolation and prevailing beliefs within the Incelverse. A recent study on user characteristics in Incel forums analyzed users from three major Incel platforms using network analysis and community detection to determine their primary concerns and participation patterns. The findings suggest that users frequently interact with content related to mental health and relationships and show activity in other forums with hateful content (Stijelja and Mishara, 2023). Similarly, Pelzer et al. (2021) investigated the spread of toxic language across different incel platforms, revealing that the engagement with toxic language is associated with different subgroups or ideologies within the Incel communities. However, these studies have generally focused on smaller subsets of users and have not examined user behavior across the entirety of the *incels.is* forum. This gap in research is noteworthy, especially when broader studies indicate that content from hateful users tends to spread more quickly and reach a larger audience than non-hateful users (Mathew et al., 2019).

3 Categorizing violent language with language models

Effectively approaching harmful language requires a nuanced understanding of the diverse forms it takes online, encompassing elements such as "abusive language," "hate speech," and "toxic language," (Nobata et al., 2016; Schmidt and Wiegand, 2017). Due to their overlapping characteristics and varying degrees of subtlety and intensity, distinguishing between these types of content poses a significant challenge. In addressing this complexity, Davidson et al. (2017) define hate speech as "language that is used to express hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." Within the research community, this definition is further extended to include direct attacks against individuals or groups based on their race, ethnicity, or sex, which may manifest as offensive and toxic language (Salminen et al., 2020).

While hate speech has established itself as a comprehensive category to describe harmful language online, the landscape of hateful language phenomena spans a broad spectrum. Current research frequently focuses on specific subfields, e.g., toxic language, resulting in a fragmented picture marked by a diversity of definitions (Waseem et al., 2017; Caselli et al., 2020a). What unites these definitions is their reliance on verbal violence as a fundamental element in characterizing various forms of harmful language. Verbal violence, in this context, encompasses language that is inherently aggressive, demeaning, or derogatory, with the intent to inflict harm or perpetuate discrimination (Waseem et al., 2017; Soral et al., 2018; Kansok-Dusche et al., 2023). Building on this foundation, we adopt the terminology of "violent language" as it aptly encapsulates the intrinsic aggressive and harmful nature inherent in such expressions. To operationalize violent language, Waseem et al. (2017) have developed an elaborate categorization of violent language online. This categorization distinguishes between explicit and implicit violence, as well as directed and undirected forms of violence in online contexts. It will serve as the fundamental concept guiding the operationalization of violent speech in this paper (see Section 5.1). By addressing various degrees of violence, this concept encompasses language employed to offend, threaten, or explicitly indicate an intention to inflict emotional or physical harm upon an individual or group.

Supervised classification algorithms have proven successful in detecting hateful language in online posts. Transformer-based models like HateBERT, designed to find such language, have outperformed general BERT versions in English (Caselli et al., 2020a). While HateBERT has proven effective in recognizing hateful language, its adaptability to diverse datasets depends on the compatibility of annotated phenomena. Additionally, although these models exhibit proficiency in discovering broad patterns of hateful language, they are limited in discerning specific layers or categories, such as explicit or implicit forms of violence. Ultimately, the capability of BERT-based models to identify nuanced patterns of hateful language, including explicit and implicit forms, depends on the dataset used for fine-tuning.

Large Language Models (LLMs) present a promising alternative in scenarios where an evaluated, labeled dataset is unavailable. Recent research has found that using LLMs, particularly OpenAI's GPT variants, to augment small labeled datasets with synthetic data is effective in low-resource settings and for identifying rare classes (Møller et al., 2023). Further, Gilardi et al. (2023) found that GPT-3.5 outperforms crowd workers over a range of annotation tasks, demonstrating the potential of LLMs to drastically increase the efficiency of text classification. The efficacy of employing GPT-3.5 for text annotation, particularly in violent language, has been substantiated, revealing a robust accuracy of 80% compared to crowd workers in identifying harmful language online (Li et al., 2023b). Even in more challenging annotation tasks, like detecting implicit hate, GPT-3.5 demonstrated a commendable accuracy by correctly classifying up to 80% of the provided samples (Huang et al., 2023). Specifically for identifying misogynistic language, Morbidoni and Sarra (2023) found that GPT-3.5 outperformed supervised baselines. While these results showcase the effectiveness of GPT-3.5 in-text annotation, there remains room for improvement, particularly in evaluating prompts and addressing the inherent challenges associated with establishing a definitive ground truth in complex classification tasks like violent language classification (Li et al., 2023b).

Although smaller, fine-tuned, discriminative language models have shown superior performance in many cases (Abdurahman et al., 2023; Kocoń et al., 2023; Mu et al., 2023; Rathje et al., 2023), LLMs stand out for their adaptability across varied tasks and their capacity to incorporate context-specific information without additional training. Their ability to generate relevant insights without requiring highly specialized datasets offers a distinct advantage, bridging the gap in research contexts with limited data resources (Huang et al., 2023; Kocoń et al., 2023; Liu et al., 2023). Given the reduced technical complexity of making API calls compared to training a BERT model, LLMs may further provide enhanced accessibility for researchers across various disciplines, making data annotation more efficient and accessible (Li et al., 2023b).

4 Summary and study outline

The Incel community has become a subject of growing academic interest due to its complex interplay of extreme views and connections to real-world violence over the last few years. While previous research has illuminated linguistic and ideological dimensions of violent language in online forums, a forum-wide analysis based on different violence categories remains lacking. By further including the user level, this study makes it possible to distinguish between the overall evolution of violent speech prevalence within the forum and observe how the prevalence of violent content shifts for individual users over their active periods in the forum. Using manual annotation in conjunction with GPT-4 for this task offers a cost-effective and flexible approach, given its pre-trained capabilities for understanding a wide range of textual nuances. By classifying different categories of violent speech, we aim to determine whether various forms of violence exhibit differing levels of prevalence within the forum and if they evolve differently over time. Results can be used to assess the threat of violence in Incel forums and help tailor intervention strategies and content moderation to the specific nature of the content, enhancing the effectiveness of efforts to mitigate harm and promote safety within online communities.

5 Materials and methods

Besides *incels.is*, platforms like *looksmax.org* and Incel-focused subreddits are key communication channels for the Incel community. After Reddit officially banned the biggest Incel subreddit *r/incel* for inciting violence against women (Hauser, 2017; Ribeiro et al., 2021b), many users migrated to alternative platforms. With a self-proclaimed 22,000 members and over 10 million posts,[1] *incels.is* has become the leading Incel forum, making it an essential resource for understanding the community.

We scraped all publically available threads from *incels.is*, yielding over 400*k* threads with more than 6.9*M* posts. These were generated by 11,774 distinct users.[2] The web scraping was performed in May 2023. We collected the raw HTML responses from the website, focusing solely on text-based content and disregarding all non-text forms of media, primarily images, which were present in ∼6.3% of posts. Most of the media content was consistent with the posts, serving as supporting references. These included memes and short clips that reinforced the points made within the posts. Given the complexity of conducting a multimodal analysis, especially regarding the assessment of violence within memes, and our specific focus on directly expressed violent language in the text, we opted not to include such media content.

Next, we employed a three-step approach, leveraging the GPT-3.5[3] and GPT-4[4] APIs. A low temperature of 0.1 for both GPT-3.5 and GPT-4, which controls the randomness of the model's output, was chosen to ensure consistent and reliable responses while maintaining the model's creativity and flexibility (Jin et al., 2023). Note that a temperature above zero does not equate to non-deterministic behavior, as OpenAI now allows for seeded randomness in their models. Following a round of manual annotation of a random sample of 3,028 posts, we iterated prompts and the number of posts per query (batch size) for both models to align their classification of violent language with the human baseline. See Section 5.2 for more detail on the content of each prompt and their iterations. Finally, we used the best-performing prompt to classify an additional 45,611 posts, which we then analyzed for temporal patterns.[5]

5.1 Categories of violence

For categorizing different types of violent language, we used a slightly adapted version of Waseem et al. (2017)'s typology of abusive language. To bridge the challenges of navigating through the variety of definitions of hate speech, Waseem et al. (2017) have identified mutual characteristics that combine previous classifications of harmful content. This makes their typology a valid reference point when classifying violent language in online forums. This concept encompasses expressions that offend, threaten, or insult specific individuals or groups based on attributes such as race, ethnicity, or gender. It extends to language indicating potential physical or emotional harm directed at these individuals or groups. Additionally, differentiating between different types of violence (explicit vs. implicit and general vs. directed) helps gain a more nuanced picture of how violence manifests online. Following this classification scheme, we distinguish violent posts between explicitly and implicitly violent, as well as between directed, undirected/general, and self-directed violence. Each post is assigned an explicit/implicit and a directed/undirected/self-directed label. Table 1 provides examples for each category.

In the context of this classification framework, explicit violent language is a very straightforward and usually directly recognizable form of violence, e.g., racist or homophobic threats. While such language can vary in context, its harmful intent is generally unambiguous. Implicit violent language is subtler and more challenging to detect. It may involve ambiguous terms or sarcasm and lacks prominent hateful words, making it difficult for human annotators and machine learning algorithms to identify (cf., Caselli et al., 2020b for a similar distinction between explicit and implicit hate speech). On the second dimension, directed violent language refers to posts that target a specific individual, either within the forum or outside. General violent language, on the other hand, addresses a group of individuals. In the Incel context, for example, this type of language is often addressed toward women or a specific ethnic group. In our analysis, we focused solely on analyzing the textual content of posts without further differentiating between violent language targeted at particular genders or forum members.

[1] These numbers are extracted from the landing page and could not be reproduced in our attempts. Out of the 22,000 users, only 11,774 appear to have engaged by posting content.

[2] This includes 890 delete users. Once a user deletes their profile, the forum replaces all occurrences of their username with *Deleted User [XXX]* but retains the now anonymous posts.

[3] `gpt-3.5-turbo-0301` at temperature 0.1.

[4] `gpt-4-1106-preview` at temperature 0.1.

[5] Our final dataset, which includes 48,639 annotated Incel posts, as well as the code to replicate our findings, can be accessed at https://osf.io/ugtz8/.

TABLE 1 Classification examples for each category.

Category	Example
Non-violent	*Pleasure has become my main purpose of getting new hobbies, music mainly is maintaining me with life*
Explicit, directed	*I hope the whore gets raped then she can press actual sexual assault charges*
Explicit, general	*Cliquey, superficial western women deserve the rope, along with the Jews that made them this way*
Explicit, self-directed	*I'm so ugly I should be killed*
Implicit, directed	*He looks like he just got back from Auschwitz*
Implicit, general	*If only women weren't like this. But females love brutality, power, and domination, so in the end they get what they deserve*
Implicit, self-directed	*The world would be better off without men like me*

5.2 Augmented classification

Based on this classification scheme, two human annotators independently labeled a subsample of 3,028 posts. Annotation was performed by one of the authors of this study (Human A) and a research assistant familiar with the field of research (Human B), both being female. They were supported by an annotation manual providing definitions and examples for each violence category, as they are presented in Section 5.1 (general description) and Table 1 (classification examples). The annotators were tasked with reviewing each comment and categorizing it accordingly. They had the option to label comments as unclear. Those comments were subsequently excluded from the baseline sample. Additionally, the research assistant could discuss any open questions or ambiguous comments with the rest of the research team for clarification. By involving multiple annotators to establish a human baseline, we ensure a robust assessment of inter-coder consistency, enabling reliable comparisons with the models' annotations. We report Cohen's Kappa (κ) (Cohen, 1968) for intercoder reliability, as it accounts for chance agreement and adjusts for imbalanced data distributions. We also report weighted and macro F1 scores to assess the performance of the classification against the human baseline. The weighted F1 score differentiates between ground truth and predicted labels, making it a suitable metric for comparing the performance of the models against the human annotators. The macro F1 score, on the other hand, is an appropriate metric for inspecting the performance regarding underrepresented classes, as it computes the F1 score for each class individually and then takes the average of those scores. We used the manually annotated sample of 3,028 posts to evaluate the performance of different query prompts and batch sizes for both GPT-3.5 and GPT-4.

We started with a basic prompt employing role-prompting, a fundamental method in prompt engineering. Assigning the model a specific role, such as an expert, has been proven to be particularly effective in guiding the model's responses (Chen et al., 2023). In our prompts, we assigned the model the role of a "moderator of an online forum, aiming to moderate abusive and hateful language." The initial prompt only included information on our classification scheme, i.e., the categories of violence. Following best practices in prompt engineering (Chen et al., 2023; Liu et al., 2023; Hu et al., 2024), we successively added additional information and instructions to the prompt. Mu et al. (2023) demonstrated that enhancing GPT-3.5 prompts with task and label descriptions notably boosts its performance. In our case, including contextual information, specifically about the posts originating from an Incel forum, significantly improved the model's performance. To further improve the prompt, we kept looking at posts where the model's classification differed from the manual annotation and tried to find patterns in the misclassifications. Further, we used a form of self-instruction, presenting those misclassifications to the model itself and asking it for advice on improving the prompt. Finally, we included instructions to explain the reasoning behind the decision in the prompt, usually in the form of the most important words. The model must produce these hints before generating the label to ensure the model focuses on the right parts of the text and avoids *post-hoc* rationalization. The instruction to provide reasons is part of all final queries, which we provide in our OSF repository created for this study (see above).

GPT-3.5 allows for a maximum of 4*k* tokens for input and output, which can contain multiple messages with different roles, such as system and user messages. The LLM treats the system message as the central reference point for its behavior, while the user message is part of the ongoing conversation. Hence, we provide the task description and classification scheme in the system message and post them in the user message. GPT-4 has a context window of 128*k* tokens. Batching multiple posts into a single classification request made the speed and cost of the classification process manageable. Otherwise, reiterating the same system prompt for each post would substantially inflate the required number of tokens. We experimented with different batch sizes, ranging from 10 to 200 posts per batch.

In practice, each classification batch looked like

```
[System Message]
<Prompt>
The posts are:
```

followed by the batch of posts

```
[User Message]
Post 1: <Post 1>
Post 2: <Post 2>
...
```

GPT-4[6] introduces a novel JSON output mode, enabling the model to generate outputs in a JSON object format instead of

[6] Since we conducted this study, OpenAI has also released a version of GPT-3.5, which supports guaranteed JSON outputs.

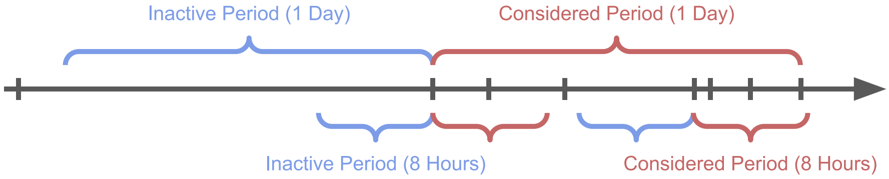
Inactive Period (1 Day)
Considered Period (1 Day)
Inactive Period (8 Hours)
Considered Period (8 Hours)

FIGURE 1
An exemplary timeline of user activity, ticks indicate posts. The user is 1-day-inactive once, while they are 8-h-inactive twice (blue). After being inactive for a given period, we observe the user's behavior for the same length of time (red).

plain text. The prompt must specify JSON schema. Our findings indicate that this mode does not alter the model's performance but significantly simplifies parsing its outputs. We used this mode for all our final classifications. Regarding data preprocessing, we limited our intervention to consolidating multiple new lines into one line. We found the model could handle the posts' raw text very well. Notably, it did not miss or confuse any post at any time. After iterating over the queries, we chose the one that performed best against the human baseline to annotate another 45,611 posts.

5.3 Time-based patterns of violent user posts

Figure 1 illustrates our method for distinguishing between active and inactive periods for individual users. We classify users as inactive if they have not made a post for at least T (e.g., 1 h, 1 day). Upon their return, we observe their behavior for the same duration, T, which we term a *session*. Posts can belong to multiple sessions since being inactive for 1 day inherently includes being inactive for 1 h, but not vice versa. This approach enables us to analyze the impact of inactivity on the prevalence of violent language in posts. To detect activity, we consider all posts, including unlabeled ones. As we cannot access viewing behavior, we need to limit our analysis of user activity to posting behavior. We repeat the following procedure for session lengths T of 1 h, 6 h, 12 h, 1 day, 1 week, 2 weeks, 30 days, and 180 days. The choice of these timespans allows us to capture the short-term, medium-term, and long-term effects of inactivity on the prevalence of violent language in posts. Due to small sample sizes, we do not report results for $T \geq 365$ days.

For each session length, we aggregate all annotated posts by their relative time since the user's first post of the session. We then divide the data into 12 equally sized bins and calculate the share of each category in each bin. To identify statistically significant trends in the prevalence of violent language, we conduct a χ^2 trend test on the resulting multinomial distribution over time for each timespan. The null hypothesis assumes no variation in the usage of violent language over time, and significance is evaluated against this assumption. To account for multiple testing, we apply Bonferroni correction, dividing the significance level by the number of tests performed (10 in our case). Significance levels are reported as $\hat{p} < 0.05$, ** indicating $\hat{p} < 0.01$, and *** indicating $\hat{p} < 0.001$ for the corrected significance levels $\hat{p}$ of the χ^2 trend tests.

To describe the trend direction, if any, we perform an ordinary least squares linear regression for each timespan, using the share of violent posts as the dependent variable and the time since the user's first post of the session as the independent variable. If inactivity reduces the prevalence of violent language, we would expect a statistically significant trend and a positive coefficient, indicating an increase in violent posts following a period of inactivity. Although the data suggests a multi-level model with random effects for users, with an average of four annotated posts per user, it is too sparse to estimate such a model reliably. Therefore, we rely on linear regression results instead.

6 Results

6.1 Performance of automated classification

Table 2 shows the pairwise Cohen's Kappa and weighted/macro F1 scores of all relevant annotation methods. Human A and B indicate the two human annotators, while GPT-3.5 presents the best-performing GPT-3.5 query and batch-size combination. GPT-4/X showcases the performance of GPT-4 with batch-size X for the best-performing query, each. Since the instruction to provide reasons for the models' decisions improved the results, it is part of all final queries.

GPT-3.5 is outperformed by GPT-4 in all metrics when comparing its labels against both human annotators. The rest of the analysis hence focuses on the performance of the different GPT-4 variants. The inter-annotator agreement between Human A and Human B, as measured by Cohen's Kappa (κ), is 0.69, indicating a substantial level of agreement. Their weighted and macro F1 scores of 0.85 and 0.77, respectively, illustrate apt performance with distinct yet varying levels of precision and recall in their annotations. Overall, Human A is less likely to label a post as violent than Human B, with 66% of posts labeled as violent by Human A, compared to 75% by Human B.

The analysis of different batch sizes reveals notable variations in the performance of GPT-4. Batch size 20 shows the highest agreement with Human A, as evidenced by its superior performance metrics. Conversely, batch size 100 aligns more closely with Human B, particularly regarding κ and weighted F1 scores. For the macro F1 score, batch size 50 exhibits the best alignment with Human B. The achieved Kappa values of 0.54 against Human A and 0.62 against Human B indicate moderate to

TABLE 2 Cohen's Kappa/weighted F1-score/macro F1-score.

	Human A	Human B	GPT3.5	GPT4/10	GPT4/20	GPT4/50	GPT4/100	GPT4/200
Human A	–	0.69/0.85/0.77	0.40/0.70/0.52	0.53/0.74/0.63	**0.54/0.76/0.63**	0.52/0.74/0.62	0.52/0.75/0.60	0.36/0.71/0.49
Human B	0.69/0.87/0.77	–	0.39/0.75/0.54	0.58/0.79/0.67	0.55/0.79/0.65	0.61/0.83/**0.67**	**0.62/0.84**/0.67	0.40/0.77/0.52
GPT3.5	0.40/0.67/0.52	0.39/0.68/0.54	–	**0.54/0.75/0.62**	0.49/0.72/0.59	0.49/0.71/0.59	0.47/0.70/0.56	0.37/0.67/0.48
GPT4/10	0.53/0.73/0.63	0.58/0.76/0.67	0.54/0.74/0.62	–	**0.75/0.86/0.78**	0.60/0.77/0.67	0.58/0.76/0.66	0.46/0.68/0.55
GPT4/20	0.54/0.75/0.63	0.55/0.77/0.65	0.49/0.74/0.59	**0.75/0.87/0.78**	–	0.69/0.83/0.74	0.65/0.81/0.71	0.44/0.71/0.51
GPT4/50	0.52/0.77/0.62	0.61/0.82/0.67	0.49/0.76/0.59	0.60/0.80/0.67	0.69/0.85/**0.74**	–	**0.72/0.87**/0.72	0.47/0.75/0.55
GPT4/100	0.52/0.78/0.60	0.62/0.84/0.67	0.47/0.77/0.56	0.58/0.80/0.66	0.65/0.84/0.71	**0.72/0.88/0.72**	–	0.51/0.80/0.59
GPT4/200	0.36/0.77/0.49	0.40/0.81/0.52	0.37/0.79/0.48	0.46/0.79/0.55	0.44/0.80/0.51	0.47/0.82/0.55	**0.51/0.83/0.59**	–

Bold numbers indicate the best performance per row, excluding humans. For the F1-scores, left indicates the ground truth, while top indicates predictions.

substantial agreement. They are similar to scores observed in other studies with comparable tasks (e.g., Haddad et al., 2019, although the authors achieved a higher agreement in one of three pairs of annotators). Macro and weighted F1 scores of 0.63 and 0.76 against Human A and 0.67 and 0.84 against Human B, respectively, indicate a high level of precision and recall in the classification of all three categories. Our weighted F1 scores of $\sim$ 0.8 align with those reported by other studies on the detection of violent language with GPT-3.5and GPT-4 (Huang et al., 2023; Li et al., 2023b). Very similar results hold for directed, undirected, and self-directed violence, which we do not report here for brevity.

Table 3 elucidates the overall label distribution across varying batch sizes, in which we observe a statistically significant shift. With increasing batch sizes, there is a discernible trend of fewer posts being classified as explicitly or implicitly violent and more as non-violent. This trend is more pronounced in the classification of implicit violence. Using a batch size of 10, 14% of all posts were labeled as implicitly violent. At batch size 200, this drops by 84%–2% of the total posts. The share of posts labeled as explicitly violent only decreases by 43% from 28 to 16%.

The label distribution generated at batch size 50 most closely aligns with the average distribution generated by the human annotators, suggesting an optimal batch size for achieving a human-like understanding of content classification. We further investigated the correlation between a post's position in a batch and its likelihood of being labeled violent. Posts positioned later in the batch were less frequently tagged as violent for larger batch sizes. This trend was consistent across different batch sizes but did not reach statistical significance. Due to the high level of agreement with humans A and B and the match in the overall class distribution, we used the labels generated by GPT-4 with batch size 50 for the remainder of our analysis.

6.2 Time-based patterns of violent user posts

Our results show that posts containing violent language, whether explicit or implicit, constitute 21.91% of all posts. 18.12% of posts contain explicit violent language, while implicit violent language accounts for 3.79% of forum posts. This leaves 78.09% of forum posts non-violent. The user analysis reveals a wide range of engagement levels. While an average of 586 posts per user appears substantial, a median of 24 posts per user indicates a very skewed distribution. About 10% of users maintained forum activity for at least 2.5 years at the time of scraping, highlighting their sustained engagement. Approximately 23.8% of forum users contributed only one post, underscoring the presence of occasional contributors within the platform's user community, while the 10% most active users have posted at least $1,152$ times. These findings underscore the diverse spectrum of user activity within the platform, ranging from highly engaged, long-term participants to sporadic contributors with limited involvement.

Figures 2A–H illustrates the temporal evolution of violent language in posts, with different time intervals as predictors for the prevalence of each violence category. Significance refers to the χ^2 tests for trends in proportions for each time interval. Regression lines are added to illustrate the overall trend. Our results indicate that within the 5 years since the forum's creation and our data collection (Figure 2A), violent language has been slightly increasing overall on a statistically significant level ($\beta = 0.006$ for explicit violence and $\beta = 0.0005$ for implicit violence). Plotting violence against time since the first post (Figure 2B), this trend is not reproduced. We find that the share of violent content remains relatively stable ($\beta = 0.0004$ for implicit violent language), with no significant changes over multiple years.

Figures 2C–J explore the impact of temporary inactivity on the prevalence of violent language. Each figure follows users for a period T, as indicated in the subfigures. The tracking takes place after these specific users have remained inactive for at least the same designated period. While we compute inactivity on the entire dataset, the plots only show annotated posts. From these figures, we observe varying results. We do not observe any statistically significant change in violent language for the 1-day (Figure 2F) and 180-day (Figure 2J) intervals. Within all other intervals, however, we observe a slight but significant increase in violent language overall, accompanied by a decrease in non-violent language. This trend is most prominent for the 2-week interval (Figure 2H) ($\beta = 0.01$ for explicit violence) and least pronounced for the 12-h window (Figure 2E) ($\beta = 0.004$ for explicit violence).

Figure 3 showcases the same analysis for the different categories of directedness. Since they do not contain any statistically relevant results, indicating that no substantial change in directed, general, or self-directed violence can be observed within the examined time

TABLE 3 Class distribution for different batch sizes *s*.

	s = 10	s = 20	s = 50	s = 100	s = 200	H-∅
Non-Violent	0.58 (1.00)	0.62 (1.07)	0.70 (1.20)	0.72 (1.24)	0.82 (1.41)	0.70 (1.21)
Explicit	0.28 (1.00)	0.26 (0.96)	0.21 (0.78)	0.22 (0.80)	0.16 (0.57)	0.22 (0.81)
Implicit	0.14 (1.00)	0.12 (0.81)	0.09 (0.61)	0.06 (0.41)	0.02 (0.16)	0.07 (0.52)

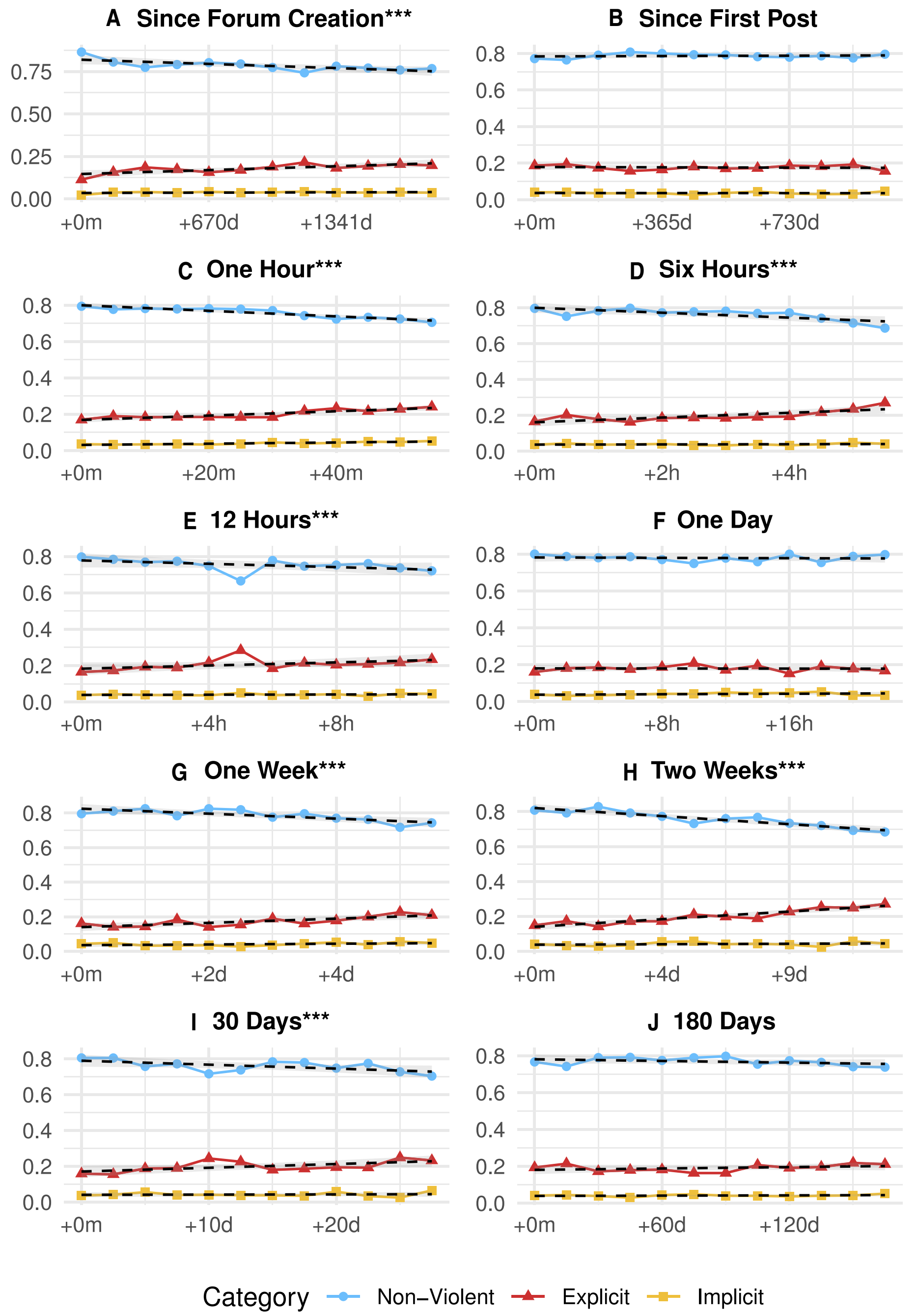

FIGURE 2
Evolution of the share of violent posts within the forum (non-violent vs. explicitly violent vs. implicitly violent). Each subfigure presents the results of separate trend tests for each time span, with the share of posts as the dependent variable and the elapsed time as the independent variable. Asterisks indicate the corrected level of significance (***$p < 0.001$). Linear regressions, represented by dashed lines, illustrate the overall trend. **(A)** Since Forum Creation***. **(B)** Since First Post. **(C)** One hour***. **(D)** Six hours***. **(E)** 12 hours***. **(F)** One day. **(G)** One week***. **(H)** Two week***. **(I)** 30 days***. **(J)** 180 days.

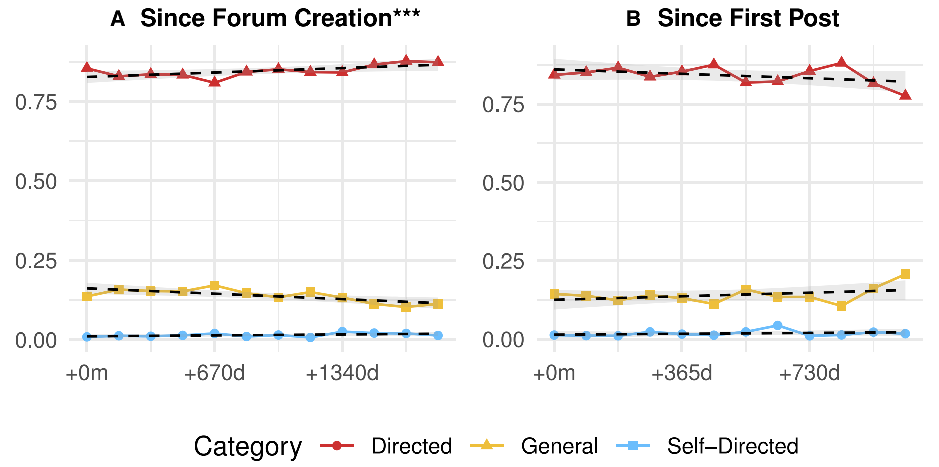

FIGURE 3
Evolution of the share of Violent Posts within the Forum (Directedness). Each subfigure presents the results of separate trend tests for each time span, with the share of posts as the dependent variable and the elapsed time as the independent variable. Asterisks indicate the corrected level of significance (***$p < 0.001$). Linear regressions, represented by dashed lines, illustrate the overall trend. **(A)** Since Forum Creation***. **(B)** Since First Post.

frames. Figure 3A reveals that the share of directed (i.e., targeted) violence increases significantly over time within the overall forum ($\beta = 0.004$). This is accompanied by a decrease in non-directed (general) violence ($\beta = -0.004$). Only considering aggregated user behavior for the time since the first post (Figure 3B), this trend appears reversed, with a slight decrease in directed violent language and an increase in general violent language. These changes, however, are not statistically significant. In this particular case, we also observe more variability in the share of violent content over time, making it harder to detect a pronounced trend. The share of self-harm content remains stable over time for both the forum and individual users (both $\beta = 0.001$).

6.3 GPT cost and speed

For the scope of this study, we spend a total of $\sim$ \$66 for OpenAI's APIs, including many iterations over all the human-annotated posts and the additionally annotated posts. Overall, we estimate GPT-3.5 and GPT-4 annotated $\sim$ 120,000 posts, including prompt iteration and batch-size experiments, which amounts to $\sim$ \$0.0005 per annotated post.

A key component of keeping the cost low is proper input batching. Our prompts are around 500 tokens long, whereas the average post is around 50 tokens long. Naively sending each post individually would have cost $550T \times \frac{\$0.01}{1{,}000T} = \$0.0055$ per post, or $\sim$ \$260 for the final set of 45,611 annotated posts. Increasing the batch-size to 50 yields a cost per batch of $3,000T \times \frac{\$0.01}{1{,}000T} =$ \$0.03, or \$20 for the final set of 45,611 annotated posts. GPT-3.5 is significantly cheaper.

The average time for GPT-4 to annotate a single post was 1 s at batch size 50. The total time for GPT-4 to annotate 120,000 posts was $\sim$ 33 h. At the time of writing, OpenAI employs strong rate limiting on their APIs, preventing us from speeding up the process by running multiple instances in parallel, rendering time constraints the more limiting factor than cost. On multiple occasions, we experienced significant slow-downs in the APIs' response time, which are confirmed by OpenAI.[7] Moving our long-running jobs to the early European morning significantly improved the experience of working with the API.

7 Discussion

Our findings reveal that 21.91% of all posts feature violent language, either explicit or implicit. We detect a subtle but statistically significant increase in overall violence on *incels.is* within the forum. The same trend is found to be more pronounced in user activity for particular time intervals, particularly in user engagement within the 2-week period. Additionally, directed violence increases over time, while self-harm consistently remains very low within the forum. This shift implies a change in the type of aggression within the community, where users resort to more targeted hostility. While these trends are very subtle, they could be explained by evolving community norms, which become more tolerant toward specific forms of violent content over time, user familiarity, or moderation effects (Gibson, 2019). Our observations align with findings from other research indicating an increase of misogynistic content and violent attitudes within Incel communities (Farrell et al., 2019) and a general rise in hate speech across various online spaces (Laub, 2019; Zannettou et al., 2020; Peters, 2022). With 21.91% of the posts exhibiting violent language, it is crucial to recognize the substantial presence of violence within these forums, emphasizing the imperative to closely monitor such platforms and contemplate legislative actions, such as implementing stricter regulations on online hate speech and harassment.

[7] https://status.openai.com

7.1 Classifying violent language with GPT

Our study indicates that LLMs can produce a sensible starting point for the zero- and few-shot classification of violent content, providing a solid foundation for further analyses. Instructing the model to identify keywords that underpin its decisions has been particularly helpful, improving its accuracy and providing a valuable reference point for a more informed comparison with human evaluators. This strategy offered a transparent framework for comprehending the model's logic, serving as a neutral benchmark for evaluating its decision-making process. However, its performance was not assessed against a standardized corpus. Other models, such as HateBERT (Caselli et al., 2020a), may perform better on datasets they are fine-tuned on. Despite this, it's important to recognize that models specialized in hate speech, including HateBERT, face difficulties in accurately classifying varied forms of violent content (Poletto et al., 2021; Yin and Zubiaga, 2021). Additionally, these models may not be explicitly designed to differentiate within distinct categories of violent language, introducing an additional layer of complexity to the classification process. Given the subtle increase in the context of a wider rise in online violent language and the large size of our dataset, which might lead to artificial effects, we must interpret these trends with caution.

The difficulty in detecting certain kinds of violent language differs significantly between categories. While explicit acts of violence, such as physical assault or overt verbal abuse, may be easier to detect through keywords or contextual cues, implicit violence often manifests in more nuanced ways that are hard even for humans to identify (Strathern and Pfeffer, 2023). These include coded language that carries a threatening subtext. For instance, users often refer to Elliot Rodger, who committed an Incel-related attack in 2014, stating posts like "Just go ER." Also, Incel-specific language is frequently inherently derogative toward women, calling them *foids*, short for feminine humanoids, and uses racist slang, e.g., *Currycel* for an Indian Incel. Herein lies an apparent strength of LLMs, which proved to be very effective at finding and classifying these Incel-specific terms. Having been trained on large parts of the internet, it is very probable that the model has encountered these terms before and learned to associate them with violence. Although misclassifications may have occurred, particularly given the challenges inherent in detecting violence of this nature, their potential impact on our work is expected to be minimal. This is because our primary emphasis is on analyzing broad trends within the platform, which means that occasional inaccuracies in classification do not impact our analysis substantially.

While the change in sensitivity for different batch sizes might seem discerning at first, it also serves as a tuneable hyperparameter. We found that manipulating the model's overall sensitivity by altering the query instead of sensitivity toward a specific class is challenging during query optimization. The batch size allows us to adjust the sensitivity to match the overall label distribution of the human annotators. It is worth noting that this adjustment substantially impacts the model's speed and cost, as discussed in Section 6.3. While other authors find similar behavior, e.g., Li et al. (2023a), we did not find research primarily focusing on this particular aspect of prompt engineering and believe a more thorough investigation could be beneficial.

The substantial agreement between GPT-4 and human annotators, alongside its accessibility and cost-effectiveness, make GPT-4 a viable alternative to traditional embedding-based classification models. Our human annotator agreement scores are comparable to those reported in prior research (Haddad et al., 2019), underscoring the challenge of attaining a Cohen's Kappa score above 0.8. Still, our agreement might be influenced by methodological limitations within the annotation process. The study relied on just two annotators, potentially skewing the analysis due to the subjective nature of detecting violent content, especially regarding more complex categories. This limitation, though resulting from practical constraints, points to an opportunity for improvement. Expanding to a broader and more diverse pool of annotators could mitigate interpretation variances and enhance classification reliability, possibly employing majority voting to achieve more balanced and unbiased results.

This study emphasizes the effectiveness of leveraging LLMs, specifically GPT-4, as annotators in intricate classification tasks, especially in identifying different types of violent content in online communities—an inherently challenging task for human annotators. By providing reasons for its classification, GPT-4 can drastically streamline situations where human annotators are uncertain. While our results provide a baseline, further research is needed to evaluate the performance of GPT-4 compared to other hate-speech-focused models. Moreover, employing LLMs, such as GPT-4, to augment the annotated sample offers distinct advantages, as it spares human annotators from the potential emotional distress of reading content containing violence against specific individuals or groups.

7.2 Violence trends within the Incel community

The results of our study align with previous research focused on radicalization within the Incel community. As noted by Habib et al. (2022), users who become part of online Incel communities exhibit a 24% increase in submitting toxic content online and a 19% increase in the use of angry language. The authors conclude that Incel communities have evolved into platforms that emphasize expressing anger and hatred, particularly toward women. In the context of online discussions on conspiracy theories, Phadke et al. (2022) modeled various radicalization phases for Reddit users, identifying different stages in radicalization that could also be applied to the Incel context in future studies.

The analyses for the 1-day (Figure 2F) and the 180-day interval (Figure 2J), as well as the period that captures the overall time since the first post on an aggregated user level (Figure 2B), do not show any statistically significant changes over time. Particularly for the longer time intervals capturing more than a month, the forum's overall increase in violent language can thus not be reproduced. However, for shorter time periods of less than a month (e.g., 1 h, 6 h, 12 h, 1 week, and 2 weeks), the increase is significant, indicating that violent language tends to spike over shorter intervals. While the 1-day interval might initially appear as an anomaly, the deviation could result from chance or other factors not accounted for in the current analysis. Therefore, it might be valuable to validate these findings with additional data to determine the reliability of this particular observation. Future research could also benefit from advanced time-series analyses to uncover deeper insights into specific trends or events within the forum.

Our findings highlight the complex relationship between user engagement duration and violent content generation. Further research may be needed to explore the underlying motivations and dynamics driving these temporal patterns in online Incel discussions. Exploring broader time-related factors, including the potential impact of COVID-19-related dynamics on online behavior—especially relevant as the pandemic overlaps with our analysis of posts from the past 5 years—holds significant importance. This consideration stems from previous studies suggesting that the pandemic contributed to shifts in behavioral patterns, leading to increased radicalization across various online forums, including those associated with Incel communities (Davies et al., 2021). Additional (computational) studies and in-person surveys with community members could provide deeper insights and guide interventions to foster more positive interactions within the forum.

Additionally, individual beliefs and attitudes of users, including their affiliation with specific subgroups within the Incel community that vary in extremism, could correlate with observed trends. It is plausible that belonging to a particular ideological subgroup may influence how members express violent content. These ideologies may affect the time spent online, the duration of active online engagement, and the posting frequency, making them relevant factors to consider in this context. It might be fruitful to examine whether the observed trends are more pronounced among specific subgroups within the community or whether they are evenly distributed over the user population. Although our results are too subtle to account for an actual pattern of radicalization, it might also be interesting to build upon these results and dive more deeply into the content of violent posts within specific time windows to see if phases of escalation can be identified.

Understanding the driving factors behind the increase in violent speech is essential to address and mitigate overall aggression levels within the forum. Investigating whether this generalized violence specifically targets certain groups, such as women or non-Incel men, could provide valuable insights into the dynamics of hostility within the community (Pelzer et al., 2021). In light of these findings, refining our analytical framework could enhance the precision of our results. Although Waseem et al. (2017)'s typology offers a solid starting point, an Incel-specific framework, such as the one proposed by Pelzer et al. (2021), which categorizes posts based on their targets—ranging from women and society to Incels themselves and ethnic groups—might yield more nuanced insights. Future research should consider these distinctions to better understand the variability in the direction of violent content. This is particularly pertinent given the observation that a significant portion of violent posts targets not only women but also "Chads," "normies," and society at large, suggesting a broad spectrum of animosity that extends beyond a single focal group.

In summary, our investigation into the evolution of violent speech within Incels forums and the intricate dynamics of ideology-driven aggression underscores the complexity of online radicalization. While we offer an overview of the evolution of specific subcategories of violence, the significance of temporal factors, ideological underpinnings, and community-specific behaviors in the online violence landscape necessitates further research. Our analysis has been limited to textual data, yet incorporating other forms of data, such as memes and short videos, through a multimodal analysis could enhance our insights (Gomez et al., 2020; Kiela et al., 2020; Bhandari et al., 2023; Chhabra and Vishwakarma, 2023). Despite the technical challenges associated with image recognition and determining the level of violence in these media, a multimodal approach in future research promises a more comprehensive understanding of the factors driving violent speech in digital communities.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Our data processing procedures did not involve any handling of private information. The user names obtained during the scraping process do not contain sufficient and valid information to make conclusions about online users' personal information. The same is true for posts directly cited in this paper. Despite offensive language in these posts, we have included them to enhance the clarity and understanding of our categorization for our readers. Both human annotators were informed of and aware of the potentially violent content in Incel posts before the annotation process, with the ability to decline annotation at any time. Both coders were given the chance to discuss any distressing material encountered during annotation. As discussions on the potential trauma or adverse effects experienced by annotators while dealing with hate speech become more prevalent (Kennedy et al., 2022), we have proactively provided annotators with a recommended written guide designed to aid in identifying changes in cognition and minimizing emotional risks associated with the annotation process.

Author contributions

DM: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. MS: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. NG: Writing – original draft, Writing – review & editing, Supervision. JP: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., et al. (2023). Perils and opportunities in using large language models in psychological research. *PsyArXiv*. [Preprint]. doi: 10.31234/osf.io/d695y

Baele, S. J., Brace, L., and Coan, T. G. (2021). From "Incel" to "Saint": analyzing the violent worldview behind the 2018 Toronto attack. *Terror. Political Violence* 33, 1667–1691. doi: 10.1080/09546553.2019.1638256

Bhandari, A., Shah, S. B., Thapa, S., Naseem, U., and Nasim, M. (2023). "Crisishatemm: multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 1993–2002. doi: 10.1109/CVPRW59228.2023.00193

Brace, L. (2021). *A short introduction to the involuntary celibate sub-culture*. Centre for Research and Evidence on Security Threats. Available online at: https://crestresearch.ac.uk/resources/a-short-introduction-to-the-involuntary-celibate-sub-culture/ (accessed March 10, 2024).

Broyd, J., Boniface, L., Parsons, D., Murphy, D., and Hafferty, J. D. (2023). Incels, violence and mental disorder: a narrative review with recommendations for best practice in risk assessment and clinical intervention. *BJPsych Adv*. 29, 254–264. doi: 10.1192/bja.2022.15

Caselli, T., Basile, V. Mitrović, J., and Granitzer, M. (2020a). Hatebert: retraining bert for abusive language detection in English. *arXiv* [Preprint]. arXiv:2010.12472. doi: 10.48550/arXiv.2010.12472

Caselli, T., Basile, V. Mitrović, J., Kartoziya, I., and Granitzer, M. (2020b). "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 6193–6202.

Chen, B., Zhang, Z. Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv* [preprint]. arXiv:2310.14735. doi: 10.48550/arXiv.2310.14735

Chhabra, A., and Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimed. Syst*. 29, 1203–1230. doi: 10.1007/s00530-023-01051-8

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull*. 70, 213. doi: 10.1037/h0026256

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proce. Int. AAAI Conf. Web Soc. Media* 11, 512–515. doi: 10.1609/icwsm.v11i1.14955

Davies, G., Wu, E., and Frank, R. (2021). A witch's brew of grievances: the potential effects of COVID-19 on radicalization to violent extremism. *Stud. Confl. Terror*. 46, 1–24. doi: 10.1080/1057610X.2021.1923188

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). "Exploring misogyny across the manosphere in reddit," in *Proceedings of the 10th ACM Conference on Web Science* (New York, NY: ACM), 87–96. doi: 10.1145/3292522.3326045

Gibson, A. (2019). Free speech and safe spaces: how moderation policies shape online discussion spaces. *Soc. Media Soc*. 5:2056305119832588. doi: 10.1177/2056305119832588

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv* [Preprint]. arXiv:2303.15056. doi: 10.48550/arXiv.2303.15056

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). "Exploring hate speech detection in multimodal publications," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Snowmass, CO: IEEE), 1470–1478. doi: 10.1109/WACV45572.2020.9093414

Habib, H., Srinivasan, P., and Nithyanand, R. (2022). Making a radical misogynist: how online social engagement with the manosphere influences traits of radicalization. *Proc. ACM Hum. Comput. Interact*. 6(CSCW2), 1–28. doi: 10.1145/3555551

Haddad, H., Mulki, H., and Oueslati, A. (2019). "T-hsab: a Tunisian hate speech and abusive dataset," in *International Conference on Arabic Language Processing* (Cham: Springer), 251–263. doi: 10.1007/978-3-030-32959-4_18

Hauser, C. (2017). Reddit bans 'Incel' group for inciting violence against women. *The New York Times*. Available online at: https://www.nytimes.com/2017/11/09/technology/incels-reddit-banned.html (accessed March 10, 2024).

Hoffman, B., Ware, J., and Shapiro, E. (2020). Assessing the threat of incel violence. *Stud. Confl. Terror*. 43, 565–587. doi: 10.1080/1057610X.2020.1751459

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., et al. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc*. ocad259. doi: 10.1093/jamia/ocad259

Huang, F., Kwak, H., and An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv*:2302, 07736. doi: 10.1145/3543873.3587368

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of women in the Incels.me forum: linguistic analysis and automatic detection. *J. Lang. Aggress. Conf*. 7, 240–268. doi: 10.1075/jlac.00026.jak

Jelodar, H., and Frank, R. (2021). Semantic knowledge discovery and discussion mining of Incel online community: topic modeling. [*arXiv*] [Preprint]. arXiv:2104.09586. doi: 10.48550/arXiv.2104.09586

Jin, Y., Li, D., Yong, A., Shi, J., Hao, P., Sun, F., et al. (2023). RobotGPT: robot manipulation learning from ChatGPT. *arXiv* [Preprint]. arXiv:2312.01421. doi: 10.48550/arXiv.2312.01421

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., et al. (2023). A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma Violence Abuse* 24, 2598–2615. doi: 10.1177/15248380221108070

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., et al. (2022). Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Eval*. 56, 1–30. doi: 10.1007/s10579-021-09569-x

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., et al. (2020). "The hateful memes challenge: detecting hate speech in multimodal memes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC: Curran Associates Inc), 14.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., et al. (2023). ChatGPT: jack of all trades, master of none. *Inform. Fusion* 99:101861. doi: 10.1016/j.inffus.2023.101861

Laub, Z. (2019). *Hate Speech on Social Media: Global Comparisons*. Washington, DC: Council on Foreign Relations, 7.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabsi, A.-L., Brewer, D., et al. (2009). Computational social science. *Science* 323, 721–723. doi: 10.1126/science.1167742

Li, J., Zhao, R., Yang, Y., He, Y., and Gui, L. (2023a). "OverPrompt: enhancing ChatGPT through efficient in-context learning," in *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Li, L., Fan, L., Atreja, S., and Hemphill, L. (2023b). "HOT" ChatGPT: the promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv* [Preprint]. arXiv:2304.10619. doi: 10.48550/arXiv.2304.10619

Lindsay, A. (2022). Swallowing the black pill: involuntary celibates'(Incels) anti-feminism within digital society. *Int. J. Crime Justice Soc. Democr*. 11, 210–224. doi: 10.5204/ijcjsd.2138

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., et al. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv*. 55, 1–35. doi: 10.1145/3560815

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). "Spread of hate speech in online social media," in *Proceedings of the 10th ACM Conference on Web Science* (New York, NY: ACM), 173–182. doi: 10.1145/3292522.3326034

Møller, A. G., Dalsgaard, J. A., Pera, A., and Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv*:2304, 13861.

Morbidoni, C., and Sarra, A. (2023). "Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5," in *CEUR Workshop Proceedings, Vol. 3571* (CEUR-WS), 31–43.

Mountford, J. (2018). Topic modeling the red pill. *Soc. Sci.* 7:42. doi: 10.3390/socsci7030042

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., et al. (2023). Navigating prompt complexity for zero-shot classification: a study of large language models in computational social science. *arXiv* [preprint]. arXiv:2305.14310. doi: 10.48550/arXiv.2305.14310

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web* (Geneva), 145–153. doi: 10.1145/2872427.2883062

O'Donnell, C., and Shor, E. (2022). "This is a political movement, friend": WHY "incels" support violence. *Br. J. Sociol.* 73, 336–351. doi: 10.1111/1468-4446.12923

O'Malley, R. L., Holt, K., and Holt, T. J. (2022). An exploration of the involuntary celibate (Incel) subculture online. *J. Interpers. Violence* 37, NP4981–NP5008. doi: 10.1177/0886260520959625

Pelzer, B., Kaati, L., Cohen, K., and Fernquist, J. (2021). Toxic language in online incel communities. *SN Soc. Sci.* 1, 1–22. doi: 10.1007/s43545-021-00220-8

Peters, M. A. (2022). Limiting the capacity for hate: hate speech, hate groups and the philosophy of hate. *Educ. Philos. Theory* 54, 2325–2330. doi: 10.1080/00131857.2020.1802818

Phadke, S., Samory, M., and Mitra, T. (2022). Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. *Proc. Int. AAAI Conf. Web Soc. Media* 16, 770–781. doi: 10.1609/icwsm.v16i1.19333

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Eval.* 55, 477–523. doi: 10.1007/s10579-020-09502-8

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., Van Bavel, J. J., et al. (2023). GPT is an effective tool for multilingual psychological text analysis. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/sekf5

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., and Long, S. (2021a). The evolution of the manosphere across the web. *Proc. Int. AAAI Conf. Web Soc. Media* 15, 196–207. doi: 10.1609/icwsm.v15i1.18053

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., and De Cristofaro, R. (2021b). Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proc. ACM Hum. Comput. Interact.* 5(CSCW2), 1–24. doi: 10.1145/3476057

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Hum.-Centric Comput. Inf. Sci.* 10, 1–34. doi: 10.1186/s13673-019-0205-6

Schmidt, A., and Wiegand, M. (2017). "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (Valencia), 1–10. doi: 10.18653/v1/W17-1101

Soral, W., Bilewicz, M., and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggress. Behav.* 44, 136–146. doi: 10.1002/ab.21737

Stijelja, S., and Mishara, B. L. (2023). Characteristics of Incel forum users: social network analysis and chronological posting patterns. *Stud. Conf. Terror.* 1–21. doi: 10.1080/1057610X.2023.2208892

Strathern, W., and Pfeffer, J. (2023). *Identifying Different Layers of Online Misogyny*. doi: 10.48550/arXiv.2212.00480

Texas Department of Public Safety (2020). *Texas Domestic Terrorism Threat Assessment*. Austin, TX.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: a typology of abusive language detection subtasks. *arXiv* [Preprint]. arXiv:1705.09899. doi: 10.48550/arXiv.1705.09899

Yin, W., and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.* 7:e598. doi: 10.7717/peerj-cs.598

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., and Stringhini, G. (2020). "Measuring and characterizing hate speech on news websites," in *Proceedings of the 12th ACM conference on web science* (New York, NY: ACM), 125–134. doi: 10.1145/3394231.3397902