Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

A Multidimensional Computational Analysis of Dehumanization in Incel Discourse

Naomi Baes[1], Luc Raszewski[2], Ekaterina Vylomova[2], Nick Haslam[1], Christine de Kock[2]

[1]Melbourne School of Psychological Sciences, University of Melbourne
[2]School of Computing and Information Systems, University of Melbourne
{n.baes, l.raszewski, ekaterina.vylomova, nhaslam, christine.dekock}@unimelb.edu.au

Abstract

Dehumanizing language is a core feature of hostile online communities, often serving to reinforce ideology and justify harm. Despite widespread claims that incel discourse uniquely dehumanizes women, it remains unclear whether dehumanization systematically differs across gendered targets. This study provides the first theory-informed, multidimensional analysis of gendered dehumanization in incel discourse, operationalizing five dimensions (negative evaluation, moral disgust, animalistic framing, mind denial, and agency denial) and applying them to 10.3 million posts from incel forums (2018–2024). We extend an existing framework by adding a mind denial dimension and refining the operationalization of animality and negative evaluation, thereby strengthening its theoretical grounding and empirical coverage. Across analyses, women-associated terms demonstrate a small but statistically significant overall increase in dehumanization. However, none of the individual dimensions show reliable women–men differences, and these patterns remain stable over time. Taken together, the results indicate that dehumanization functions as an enduring representational baseline in this community: both women- and men-associated terms are evaluated negatively and embedded in dehumanizing contexts. These findings illustrate how theory-informed computational measures can characterize the multidimensional ways ideological harm is expressed in online communities.

Introduction

'Incel' (a portmanteau of 'involuntary celibate') refers to an online subculture composed predominantly of men who attribute their romantic or sexual frustrations to women and adopt the label to describe a perceived inability to form romantic relationships.[1] Once dismissed as fringe, incel communities have attracted increasing scholarly and public attention due to their virulent misogyny, nihilistic worldview, and documented links to online radicalization and real-world violence (Hoffman, Ware, and Shapiro 2020; Blake and Brooks 2023). A central feature of incel discourse is its dehumanizing language, which frames women as subhuman or morally impure.

Dehumanization, perceiving others as less than fully human, is a well-established mechanism for legitimizing harm and maintaining social hierarchies (Haslam 2006; Haslam and Stratemeyer 2016). Although social psychologists have examined the attitudinal and motivational foundations of misogyny, far less attention has been paid to how dehumanization is *linguistically* constructed and reinforced within the manosphere. Meanwhile, natural language processing (NLP) research on online abuse has largely concentrated on hate speech detection (Salminen et al. 2018; Hoeken et al. 2023; Zsisku, Zubiaga, and Dubossarsky 2024), often overlooking the theoretically grounded, multidimensional mechanisms through which ideological harm is expressed.

To address these gaps, we integrate social psychological theory with large-scale NLP to analyze dehumanization in incel discourse. We adapt and extend a framework developed by Mendelsohn, Tsvetkov, and Jurafsky (2020) to operationalize five dimensions of dehumanization: negative evaluation, moral disgust, animalistic framing, mind denial, and agency denial. The latter two capture cognitive and volitional devaluation that is pervasive in misogynistic discourse yet underexamined in prior work. We apply this multidimensional framework to 10.3 million incel forum posts (2018–2024). Rather than treating dehumanization as monolithic, we examine how its distinct forms vary across targets and over time. Figure 1 illustrates how dimensions co-occur.

Source: Inceldom Discussion (March 2024)
Thread: "I tried to approach young hot foids at the pool"

*"**They're like robots**[A]. They go through life simply **using others for their own gain**[D]. **Zero empathy or compassion**[M]."*

Figure 1: Illustrative incel post showing multiple dehumanization mechanisms: [M] Mind denial; [A] Agency denial (including mechanistic/dominance framing); [D] Moral disgust.

To illustrate how these patterns become embedded in learned representations, we fine-tuned a language model on incel posts and prompted it with "Women deserve to be [MASK]." The vanilla, general-domain pretrained BERT model (Devlin et al. 2019) produced completions such as *"happy"* and *"protected"*, whereas the incel-fine-tuned model predicted *"raped"* and *"killed"*, demonstrating how domain-specific exposure shapes the model's semantic priors and normalizes extreme dehumanization.

[1]**Warning:** Dehumanizing language appears in this paper.

Our Contribution

The multidimensional framework of Mendelsohn, Tsvetkov, and Jurafsky (2020) was originally developed and applied to analyze dehumanizing language toward LGBTQ groups in news media, but its empirical use has remained limited (e.g., applications to substance-related discourse; Giorgi et al. 2023) and it has not been tested in ideologically extreme online environments. We extend and validate this theory-informed framework in incel forums, an understudied domain where claims of extreme, gender-asymmetric misogyny are widespread but rarely evaluated using grounded semantic measures. We introduce two substantive extensions: 1) adding *mind denial*, a core construct in social psychological models of dehumanization absent from the original framework; and 2) replacing lexically anchored animal metaphors with a more generalizable *animality–centroid* representation capturing semantic proximity to animal-related concepts; we also refine the negative-evaluation index. These adaptations broaden theoretical coverage and improve robustness in a domain characterized by abnormal, rapidly shifting linguistic norms (De Kock 2025). Methodologically, we implement refined operationalizations of five dimensions (negative evaluation, moral disgust, animality, mind denial, and agency denial), each measured using NLP indices aligned with social psychological theory, enabling corpus-scale quantification. Conceptually, our design provides the first multidimensional assessment of gendered dehumanization in incel discourse: although women are a salient target, men are also frequent objects of denigrating, self-directed rhetoric (Bogetić et al. 2023), making gender comparison essential. Therefore, the present study examines three research questions:

RQ1. Are women framed more *negatively* than men?

RQ2. Are women more dehumanized than men, as indexed by stronger linguistic associations with (i) *moral disgust*, (ii) *animality*, (iii) *mind denial* (lacking thoughts, feelings, intentions), and (iv) *agency denial*?

RQ3. Have these dehumanizing representations changed differently over time for women versus men?

Related Work

Dehumanization

Dehumanization has been defined as "the act of perceiving or treating people as if they are less than fully human" (Haslam and Stratemeyer 2016). Psychological theories propose that it involves explicitly or implicitly likening people to nonhuman entities. According to one prominent account (Haslam 2006), dehumanization can occur either by viewing others as animal-like, as when people are denied uniquely human qualities such as rationality, morality, and culture ("animalistic dehumanization"), or by likening them to inanimate objects, as when they are stripped of emotion, individuality, and agency ("mechanistic dehumanization").

Research on dehumanization has assessed it in many ways, from subtle denial of uniquely human characteristics to blatant objectification or use of degrading animal metaphors (Haslam et al. 2008). It often involves stripping people of the capacity to have mental states such as emotions and complex cognitions, a phenomenon known as "mind denial" (Waytz et al. 2010). Although dehumanization of a group is usually accompanied by negative evaluation, it is distinct from hatred or dislike, and even has a distinct neural signature (Cikara, Eberhardt, and Fiske 2011; Jack, Dawson, and Norr 2013; Bruneau et al. 2018).

Numerous social groups are documented targets of dehumanization, with research studies focusing on groups based on race, gender, sexuality, mental illness, disability, and social class, among others (Haslam 2006; Haslam and Loughnan 2014). Dehumanizing perceptions of groups have been shown to enable aggression and social exclusion, and to predict adverse behavior towards outgroups, to minimize in-group guilt, to justify violence, and to foster indifference to suffering and oppression (Kelman 1973; Opotow 1990; Bandura 1999).

Misogyny in Incel Discourse

Incel ideology systematically dehumanizes and sexualizes women, portraying them as biologically inferior or even as a distinct "species". Prior work documents the pervasive use of misogynistic humor, slurs (e.g., 'femoid' as a lexical replacement for 'woman'), and rhetorical strategies that trivialize and normalize gender-based hatred and violence within incel communities (Scotto di Carlo 2023; Chang 2022) and the broader manosphere (Ging 2019). Such discourse situates incels within a wider ecosystem of online misogyny that blends explicit extremism with everyday dehumanizing language. Central tenets of incel ideology — biological determinism, rigid gender hierarchies, and grievance over perceived sexual exclusion (Baele, Brace, and Ging 2024) — frame social relations as biologically fixed and hierarchical, providing a theoretical basis for persistent negative evaluation and the selective use of dehumanizing framings.

Focusing on metaphors in the Incels.is dataset, Bogetić et al. (2023) find that a majority of dehumanizing metaphors target women (52.24%), consistent with evidence that women are central targets in incel and broader manosphere discourse. Crucially, however, they also report substantial metaphorical dehumanization directed toward incels themselves (26.87%) and toward other men (20.9%). This pattern suggests that dehumanization in incel discourse is not confined to out-group targets, but is distributed across multiple social actors, motivating a within-community comparison that can reveal how different forms of dehumanization are selectively applied to in-group versus out-group targets. Complementing this work, Lacalle et al. (2024) analyze misogynistic zoomorphism in a Spanish manosphere forum and show that animal terms such as "seal," "whale," "bunny," "bitch," "vixen," "lizard," and "viper" are overwhelmingly used to refer to women in ways that foreground negatively evaluated physical, sexual, and moral traits (e.g., obesity, hypersexuality, perceived moral corruption), while positive associations with those animals are almost never invoked.

Beyond the tendency to use dehumanizing language, incel discourse is also intensely violent in nature. Baele, Brace, and Coan (2021) identify the incel online ecosystem as one of the most extreme factions of the manosphere, while also sharing with that movement a tendency to justify crimes

against women. Longitudinal analyses further indicate that violent language in incel spaces has increased steadily in recent years, with pronounced surges in both posting volume and violent rhetoric in response to the COVID-19 pandemic (Baele, Brace, and Ging 2024). Overall, this establishes incel discourse as a setting in which dehumanization, violence, and ideological reinforcement are tightly intertwined.

Computational Work

While hate speech detection has received substantial attention in NLP, dehumanization has received comparatively less focus despite often constituting an important ideological substrate underlying such speech (Engelmann, Trolle, and Hardmeier 2024). In many hate speech datasets, dehumanization appears only as an auxiliary label within broader annotation schemes. Large-scale analyses of gendered hate speech built on such resources further illustrate this limitation, as dehumanization is typically treated as one signal among many (Coppolillo 2025).

More recent computational work has moved beyond binary classification of hate labels by developing datasets and models that target specific categories of dehumanization. For example, Joshi (2024) introduced a large Reddit dataset focused on animalistic dehumanization, annotating both dehumanizing expressions and their target identities. Extending this direction, Assenmacher et al. (2025) present a theory-informed bilingual dataset covering both explicit and implicit animalistic and mechanistic dehumanization across social media. In parallel, work on extremist and misogynistic online communities shows that ideological language co-evolves with community structure, highlighting the importance of modeling socio-temporal dynamics to capture group-specific linguistic shifts (De Kock 2024, 2025).

However, existing computational approaches largely frame dehumanization as an instance-level prediction task, focusing on comment-level detection and target labeling using fine-tuned transformer classifiers and prompted LLMs evaluated on annotated datasets (e.g., Saffari et al. 2025). Although this framing is useful for identifying whether and where dehumanization occurs and for comparing model performance across groups, it is not designed to capture dehumanization as a multidimensional phenomenon. Furthermore, LLMs (e.g., GPT-4) may rely on broad, neutral classifications that overlook the culturally embedded and ideologically charged dimensions of discourse (Breazu et al. 2024). Consequently, prior work rarely examines how mechanisms such as mind denial and animalistic framing co-occur, vary across social targets, or persist over time within a single community, limiting insight into the structure and longitudinal dynamics of dehumanizing discourse.

Method

Dehumanization Framework

We extend Mendelsohn, Tsvetkov, and Jurafsky (2020) to quantify five theoretically grounded dimensions of dehumanization in incel discourse: (1) negative evaluation, (2) moral disgust, (3) animalistic framing, (4) mind denial, and (5) agency denial. Our contribution is twofold: we add *mind denial*, a central construct in psychological models of dehumanization missing from the original framework, and we refine existing indices by updating a negative-evaluation measure and replacing the animal-metaphor component with an *animality centroid* capturing semantic proximity to animal-related concepts. Each dimension is operationalized using lexicon-based or connotation-frame scores and embedding-based associations between group labels and concept centroids (Table 1). Because several indices rely on a common embedding-based procedure, we describe this pipeline before detailing the dimension-specific operationalizations.

Shared Embedding-Space Procedure Several indices rely on a shared embedding-based procedure for quantifying semantic associations between target groups and dehumanization dimensions. All measures are computed within year-specific contextual embedding spaces using cosine similarity. Unlike static embeddings (e.g., word2vec), which assign a single context-invariant vector to each word (Mendelsohn, Tsvetkov, and Jurafsky 2020), contextualized representations vary by sentential context, enabling finer-grained modeling of semantic variation. For each group, we compute a centroid by averaging contextualized token embeddings of its associated terms. For neighborhood analyses, we retrieve the 500 nearest neighbors to the centroid within the same yearly space; for affective or dominance analyses, only neighbors present in the NRC-VAD lexicon (Mohammad 2025) are retained and their scores are averaged. This procedure underlies the animality and mind-denial associations and the dominance-based agency and valence diagnostics. Embeddings are not aligned across years, so similarity is interpreted only within each yearly model. Full extraction and retrieval details are provided in Appendix section "Contextual Embedding Extraction".

Negative Evaluation of Target Group Although dehumanization is conceptually distinct from negative evaluation and cannot be reduced to hatred, the two are empirically correlated. Negative evaluation of a target group is therefore a theoretically motivated marker of dehumanizing representation (Mendelsohn, Tsvetkov, and Jurafsky 2020). In language, evaluative meaning can be captured through Valence, a primary dimension of affect (Russell 2003), and through the Evaluation (good/bad) dimension of connotational meaning (Osgood, May, and Miron 1975). We operationalize negative evaluation using three complementary measures to provide a multi-level triangulation of evaluative meaning: (i) local lexical context surrounding group labels with valence lexica, (ii) predicate-level writer perspective toward group-referent arguments using connotation frames (Sap et al. 2017), and (iii) semantic neighborhood structure in embedding space. Our primary inferences are based on the first two measures, which directly index evaluative framing in discourse, whereas the embedding-based measure captures broader semantic drift.

Valence Index For each year, we compute the count-weighted mean valence of lemmatized context words[2] oc-

[2]Using SpaCy's "en_core_web_trf" (RoBERTa-base), which achieves 98% POS-tagging accuracy.

Element	Quantification
Negative Evaluation	• Valence index (count-weighted mean valence-matched neighboring lemmas) • Perspective polarity (Connotation Frames) • Valence-matched nearest neighbors in group embedding space
Moral Disgust	• Mean cosine similarity of group label vectors to *Purity* centroid
Animality	• Mean cosine similarity of group label vectors to *Animal* centroid
Mind Denial	• Mean cosine similarity of group label vectors to *Mind Perception* centroid
Agency Denial	• Agentivity index (Connotation Frames) • Dominance-matched nearest neighbors in group embedding space

Table 1: Quantification of Dehumanization Elements.

curring within ± 5 tokens of each target term, excluding lemmas in the target lists, using psycholinguistic valence norms (Warriner, Kuperman, and Brysbaert 2013). Scores are normalized to $[0, 1]$ and averaged across terms (Appendix section "Warriner Norms Lexicon: Valence"). Originally introduced by Baes, Haslam, and Vylomova (2024) to operationalize sentiment, this index captures the mean affective valence of linguistic contexts surrounding group labels, with lower values indicating more negative, and higher values more positive, emotional associations. This replaces the paragraph-level valence scores used by Mendelsohn, Tsvetkov, and Jurafsky (2020), providing a more fine-grained indicator of negative evaluation. Consequently, it captures local evaluative framing in discourse.

Perspective Score We estimate writer perspective toward women and men using *Connotation Frames* (CF; Rashkin, Singh, and Choi 2016).[3] We first extract subject–verb–object (SVO) tuples from each post using `spaCy`'s English dependency parser. Tuples are retained if either the subject or object noun phrase (determiner–adjective–noun span) contains a target label from the feminine or masculine term lists. To ensure adequate sampling, we further restrict tuples to those whose head verb lemma appears in the CF lexicon. For each matched tuple, we map the head verb to its CF entry and record the writer's *Perspective toward the Subject* (Writer $\rightarrow$ Subject) when the target appears as the subject, and *Perspective toward the Object* (Writer $\rightarrow$ Object) when it appears as the object. If a target appears in both syntactic roles, both perspective scores are retained. To control variance and computational load, we apply reservoir sampling, capping the number of tuples per group–year at 200. The yearly CF perspective score is then computed as the mean of all retained SVO-level perspective scores for that group–year (range: $-1, 1$; higher values = more positive writer perspective). See Appendix section "Connotation Frames" for further details. Because CF perspective scores rely on predicate–argument structure, they provide an estimate of writer stance toward the target group.

Neighbor Valence in Group Embedding Space As a diagnostic complement to the context- and predicate-based measures, we assess the affective polarity of semantic neighborhoods associated with gendered group labels using the shared embedding-space procedure. For each group–year, we aggregate the valence scores of the 500 nearest neighbors that appear in the NRC-VAD lexicon (Mohammad 2025). The reported score is the mean valence of the retained neighbors, yielding a single group–year estimate [0–1]. This measure reflects the average affective polarity of concepts semantically associated with the group, rather than the evaluative tone of local linguistic contexts in which the group is mentioned. Accordingly, it captures global semantic drift in group representations and is treated as a robustness-oriented diagnostic rather than a primary index of evaluative framing.

Dehumanization Elements

Moral Disgust Disgust is a central component of dehumanization, based on the perception that members of a target group are bestial and contaminating. As an emotion related to perceived impurity or depravity, disgust is associated with seeing a group as lacking moral worth (Sherman and Haidt 2011), being animal-like (Buckels and Trapnell 2013), and is a common basis for harming, reviling, and avoiding group members (Hodson and Costello 2007). To operationalize this dimension, we construct a Moral Disgust concept vector by averaging contextualized embeddings of 46 words derived from the Purity subdictionary of the Moral Foundations Dictionary (Graham, Haidt, and Nosek 2009).[4] Starting from the 29 original stems (e.g., *disgust*, *dirt*, *pervert*, *slut*), we expand the list using their inflectional variants, filtering to retain only in-vocabulary forms. Using the shared embedding-space procedure, we compute cosine similarity between group label vectors and the Moral Disgust centroid, averaging scores by group and year, ranging from [0–1]. Higher similarity values indicate stronger semantic association with moral impurity and contamination, whereas lower values reflect weaker association. Yearly averages track how strongly each group becomes associated with moral disgust.

[3] https://maartensap.com/connotation-frames/

[4] Original Purity dictionary list available at: https://moralfoundations.org; see the expanded list used in the present study at the GitHub repository: https://github.com/lraszewski/words-against-women/blob/master/data/concepts/purity.csv

Animal Likeness Animalistic dehumanization involves attributing non-human, biologically driven, or base traits to people, casting them as less evolved or subhuman. While prior work has focused on explicit animal metaphors (e.g., likening groups to "vermin"; Steuter and Wills 2010), we operationalize animalistic dehumanization more broadly as semantic association with animal concepts. To this end, we construct an *Animal* concept centroid using a lexicon of common animal terms drawn from Wikipedia's list of animal names by species and taxon.[5] Using the shared embedding-space procedure, we estimate cosine similarity between group label vectors and the Animal centroid, aggregating scores by group-year. Values range from [0–1], with higher values indicating stronger semantic association with animal-related concepts. This embedding-based approach captures not only overt animal metaphors (e.g., *pig*, *snake*), but also subtler forms of biologically reductionist or non-agentic framing that may not involve explicit metaphorical language. This extends the vermin metaphor used by Mendelsohn, Tsvetkov, and Jurafsky (2020) and draws on prior work by Lacalle et al. (2024) who show extensive use of animal metaphors in the manosphere.

Mind Denial Dehumanization typically involves denying people attributes that are seen as uniquely human. The "mind perception" account of dehumanization (Waytz et al. 2010) argues that mental capacities constitute many of these attributes, making mind denial a key index of dehumanization. Research shows that people perceive minds along two dimensions (Gray, Gray, and Wegner 2007). The *Agency* dimension includes cognitive capacities like thinking, planning, and self-control, which distinguish adult humans from animals and young children. The *Experience* dimension includes capacities to feel emotions, pleasure, and pain, which distinguish humans and animals from inanimate objects.

To operationalize mind denial, we estimate the degree to which group labels are semantically associated with mental state language. We construct a *Mind Perception* concept centroid from 326 terms drawn from the Mind Perception Dictionary (Schweitzer and Waytz 2021), which captures cognitive, affective, and perceptual states.[6] Using the shared embedding-space procedure, we compute cosine similarity between group label vectors and the Mind Perception centroid, aggregating scores by group and year. Ranging from [0–1], higher similarity values indicate stronger attribution of mental states (i.e., greater attribution of thought, feeling, or sentience) whereas lower values reflect mind denial, an element of mechanistic dehumanization. This extends the indices in Mendelsohn, Tsvetkov, and Jurafsky (2020) to enable detection of subtle shifts in how social groups are portrayed with respect to mental agency and personhood, even in the absence of explicit dehumanizing language.

Agency Denial Dehumanization often involves perceiving members of a target group as lacking agency — that is, as passive automatons, incapable of intentional, controlled, or purposeful action (Tipler and Ruscher 2014). Agency denial specifically concerns whether a group is represented as able to act upon the world, exert control, or bring about outcomes, rather than merely experiencing mental states (Gray, Gray, and Wegner 2007). We measured lexical indicators of *denial of agency* by testing the extent to which writers attribute agency, specifically intentional action and control, to group members, using Mendelsohn, Tsvetkov, and Jurafsky's (2020) conceptualizations of agency (range: 0–1, where higher values indicate greater attributed agency).

Agentivity Index We quantified perceived agency of target groups using the Connotation Frames of Agency lexicon (Sap et al. 2017), which contains agency-annotated verb lemmas (2,081 single-word verbs and 65 multi-word verb patterns). Each verb is labeled with an agency polarity toward its subject (–1 = low agency, 0 = neutral, +1 = high agency). High-agency verbs (e.g., *attack*, *praise*) imply intentional, agentive action, whereas low-agency verbs (e.g., *need*, *doubt*) imply passivity or diminished control. Using spaCy's dependency parser, we extracted subject–verb pairs and retained only tuples in which the subject noun phrase contained a target group label. Multi-word verb patterns (e.g., *'toy with'*, *'tick off'*) were matched by lemmatizing and aligning phrases to the head verb. To limit variance across years and terms, we applied reservoir sampling (200 tuples per term–year). For each term and year, we computed a labeled-only agentivity index, defined as the proportion of tuples whose verbs carry positive agency polarity among all tuples whose verbs have any agency label. This produces an index on the 0–1 scale, where higher values indicate greater perceived agency attributed to the group. Because the index is grounded in predicate–argument structure, it captures explicit sentence-level attributions of agency rather than broader semantic associations.

Neighbor Dominance in Group Embedding Space As a complementary diagnostic for agency denial, we assess how strongly group representations are embedded in semantic regions associated with perceived power, control, and dominance. Whereas the agentivity index captures local syntactic attributions of intentional action, this embedding-based measure reflects the broader semantic positioning of groups with respect to dominance-related associations. Using the shared embedding-space procedure, we identify the 500 nearest neighbors of each group centroid within each yearly embedding space and retain only those present in the NRC-VAD lexicon (Mohammad 2025). We then compute the mean dominance score of the retained neighbors, yielding a single dominance index per group–year (0–1). Higher values indicate stronger associations with power and con-

[5] https://en.wikipedia.org/wiki/List_of_animal_names The final dictionary and preprocessing code are available in the project repository: https://github.com/lraszewski/words-against-women/blob/master/data/concepts/animals.csv

[6] Original dictionary available at https://www.shaneschweitzer.com/materials. Following common practice in social-psychological lexicons, we retain the dictionary's stem-based entries (e.g., *Believ**, *Attitud**, *Awar**), which implicitly cover their inflected surface forms. The exact term list used in our analysis is provided in the project repository: https://github.com/lraszewski/words-against-women/blob/master/data/concepts/mind-overall.csv

trol, as indexed by dominance ratings in the NRC-VAD lexicon (illustratively, high-dominance terms such as *govern*, *command*, *success*), whereas lower values reflect association with weakness, dependence, or subordination (e.g., low-dominance terms such as *frail*, *helpless*, *penniless*). Lower dominance is interpreted as greater denial of agency, reflecting a representational context in which the group is semantically situated closer to concepts of passivity or powerlessness, even when agency-denying predicates are absent.

Statistical Strategy

Analyses use term–year observations for 60 women-associated and 39 men-associated terms from 2018–2024. Because terms repeat across years, all models are estimated using Ordinary Least Squares regression with year fixed effects to absorb corpus-wide temporal shifts and standard errors clustered by term. For RQ1 (negative evaluation), we estimate the women–men contrast on each index using a one-sided test, consistent with the directional hypothesis that women are more negatively evaluated than men, and report unstandardized coefficients (B) to preserve interpretability on the original scales. For RQ2 (dehumanization profiles), we orient and standardize all dimension scores and fit a profile model[7] to evaluate both the aggregated Group effect and the Group × Dimension interaction. Because RQ2 specifies directional hypotheses ("women more dehumanized than men"), both the overall Group contrast and the dimension-specific contrasts are tested using one-sided tests (H_1: women $>$ men), with Holm–Bonferroni–corrected one-sided p-values reported for the dimensions. For RQ3 (temporal change), we estimate Group × Year interactions (treating year categorically) to assess whether women–men differences shift over time, weighting models by the number of tuples contributing to each term–year estimate when appropriate. A schematic overview of the full yearly analysis pipeline is provided in Figure 7.

Materials

Incel Dataset

We use the manosphere subset of De Kock et al. (2025), containing 10.3 million posts (228.8M tokens) from Incels.is between 2017–2024, following the forum's migration from Reddit after its ban for inciting violence against women (Nov 7, 2017; Ribeiro et al. 2021). The dataset includes posts from 16,654 users (median = 25 posts per user; 44 characters per post), along with metadata (e.g., usernames, timestamps, and forum tags). Most content comes from the "Inceldom Discussion" forum (see Table 6). Preprocessing involved extracting posts yearly and parsing the HTML to remove quoted material from replies. Figure 2 shows the final distribution. We remove 2017 for analyses.

Contextualized Embeddings

All embedding-based analyses in the present study use the same year-specific, monolingual contextual embedding pipeline, with cosine similarity interpreted only within a given year and model. To capture context-specific word usage, we extracted contextualized embeddings from year-specific fine-tuned `bert-base-uncased` models.[8] We validated the fine-tuning using a masked-token prompt diagnostic. Table 2 reports the top predictions.[9]

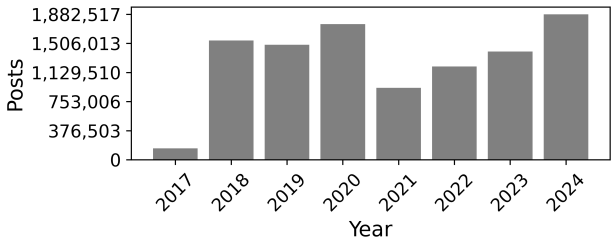

Figure 2: Posts per year (2017–2024).

Base Model		Fine-tuned Model	
Token	**Prob.**	**Token**	**Prob.**
happy	0.131	raped	0.563
protected	0.084	killed	0.050
loved	0.079	loved	0.026
respected	0.019	destroyed	0.019
equal	0.016	tortured	0.015
obeyed	0.016	dead	0.014
touched	0.014	fucked	0.013
treated	0.013	punished	0.013
married	0.012	enslaved	0.013
raped	0.012	hated	0.013

Table 2: Top 10 Predictions for the masked token in the sentence *"Women deserve to be [MASK]"*, comparing the 2024 BERT base model and incel-fine-tuned language model.

To model diachronic change, we fine-tuned a separate masked language model for each year (2018–2024) on the corresponding yearly corpus using the standard masked language modeling objective (Devlin et al. 2019) (learning rate 2×10^{-5}, weight decay = 0.01, 500 warm-up steps). Fine-tuning quality was verified through a masked-token prediction diagnostic (see Table 2). Token-level representations were then extracted at the target-word position(s) and aggregated following established procedures (Appendix section "Contextual Embedding Extraction"). In sum, we averaged subword representations (excluding special tokens) across multiple contexts and lower Transformer layers, using centered context windows (Vulić et al. 2020).

Target Term Lexica

Labels for group members were developed through consultation with a manosphere expert and corpus analysis. See Appendix section "Target Term Lexica" for more de-

[7] $z \sim$ Group $\times$ Dimension $+$ Year.

[8] https://huggingface.co/google-bert/bert-base-uncased

[9] Training scripts are available at: https://github.com/lraszewski/words-against-women.

tail about lexicon construction and the relative prominence of terms in the incel forums.

Labels for Women To examine representations of women in incel discourse, we compiled a 60-term lexicon including general references (e.g., "woman[en]," "female[s]," "girl[s]," "mother[s],"), pronominal forms (e.g., "her[s]," "she['s],"), and subculture terms (e.g., "pj[s]," "sloot[s]," "roastie[s]," "becky[ies]," "stacy[ies]," "waifu[s]").

Labels for Men For comparison, the men lexicon includes 39 terms, comprising general identifiers (e.g., "man[men]," "male[s]," "boy[s]," "guy[s]") and terms common to the manosphere (e.g., "incel[s]," "chad[s]," "gigachad[s]," "soyboy[s]," "simp[s]," "numale[s]").

Results

Semantic Neighborhoods of Men and Women

Global nearest-neighbor analyses offer a descriptive view of how women- and men-associated centroids are positioned in semantic space over time (Appendix section "Nearest-Neighbor Extraction"). Across all years, women's neighborhoods are dominated by evaluative and affective descriptors - often negative or sexualized (e.g., *snooty*, *sneering*, *lusty*, *drowsy*) - alongside judgment predicates (e.g., *dislikes*, *abuser*). Men's neighborhoods partially overlap but show more cognitive or dispositional descriptors (e.g., *brainy*, *thinker*, *aspirer*, *careerist*) alongside affective terms (e.g., *antsy*, *lusty*). These patterns align with lexicon-level trends (Appendix section "Target Term Lexica"): women-associated labels shift from generic identifiers (e.g., *girl*, *female*) toward more explicitly derogatory forms (e.g., *foid*, *whore*), whereas men-associated labels show declining use of identity-defining terms (e.g., *incel*, *chad*) and a modest increase in *normie*. Overall, descriptor types remain stable from 2018–2024, but the specific labels referring to each group change over time.

Incels Negatively Evaluate Women and Men

As shown in Figure 3, both women- and men-associated terms show increasingly negative evaluative tone from 2018–2024 on both the valence index and the perspective score. Aggregated across years, women score slightly more negatively than men on both indices (Table 3). For the valence index, the women–men contrast is small and non-significant. For the perspective index, the women–men contrast is likewise small but significant. In both cases, standardized effects are modest (≈ 0.19 SD), indicating that despite a corpus-wide shift toward greater negativity, women and men are evaluated similarly negatively in incel discourse.

Robustness analyses confirmed that these results are not driven by a small subset of highly salient or frequent labels. Leave-one-term-out refits produced stable women–men contrasts in contextual valence (range: $B = .008$–$.014$), with no sign reversals even when excluding charged slurs (e.g., *foid*, *roastie*), neutral descriptors (e.g., *woman*, *man*), or pronouns (e.g., *she*, *he*). Frequency-stratified analyses likewise yielded comparable estimates across high- and low-frequency terms. These checks indicate that the observed

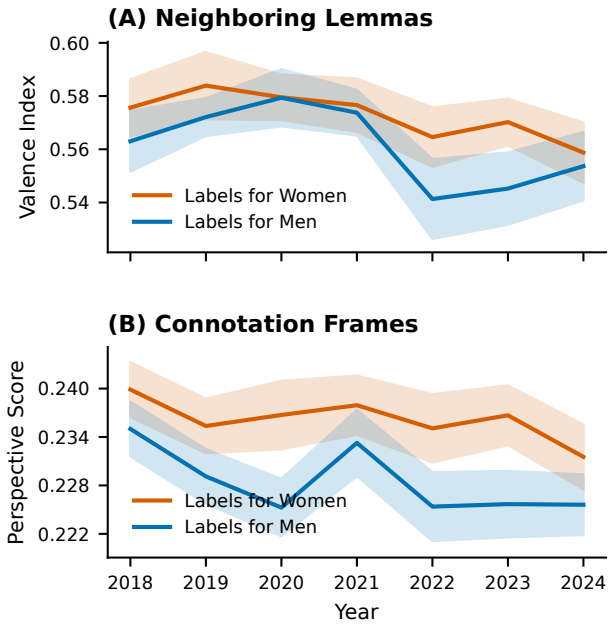

Figure 3: Negative evaluation of women- and men-associated labels in incel discourse (2018–2024). Panel (A) shows the mean contextual valence of words occurring within ± 5 tokens of each group label, averaged across terms within each group. Panel (B) shows writer perspective scores derived from Connotation Frames for subject–verb and verb–object tuples containing group labels.

Note. Panel (A) reports unweighted term–year averages within each group. Panel (B) reports pair-count–weighted grand means across terms, where weights reflect the number of extracted subject or object tuples associated with Connotation Frames verbs. Shaded ribbons indicate ± 1 standard error of the mean (SEM). Y-axes are truncated in both panels to improve visibility of small differences, and the y-axis ranges differ across panels; thus, vertical distances should not be compared directly across panels.

patterns reflect distributed properties of incel discourse rather than idiosyncratic effects of particular labels.

As a complementary diagnostic, we examined global semantic associations using NRC-VAD (Appendix section "NRC-VAD Lexica: Valence and Dominance"). Unlike the downward trend in our context-sensitive indices, centroid valence shows a modest upward drift, indicating that global semantic movement can diverge from directed evaluative stance. The two trends are correlated ($r = .77$, $p = .043$) but share only 59% variance. Because our primary measures condition on syntactic role and local context, they capture evaluative stance directly, confirming that rising negativity is a feature of incel discourse toward both genders.

Dehumanization of Both Genders

To assess whether incel discourse differentially dehumanizes women and men (RQ2), we compared their scores across four dehumanization dimensions. Women-associated

	Women	Men	Contrast	
Index	M (SD)	M (SD)	B	**95% CI**
Valence	.57 (.06)	.56 (.05)	0.0116	[−.010, .033]
Perspective	.24 (.06)	.23 (.04)	0.0096	[.000, .019]

Table 3: Descriptive statistics and women–men contrasts for negative evaluation indices (2018–2024). Contrast coefficients (B) represent unstandardized women–men differences; 95% confidence intervals are shown in the table.

One-sided p-values: Valence = .145; Perspective = .024.

terms scored slightly higher overall ($\beta = 0.10$, SE $= 0.05$, 95% CI $[-0.01, 0.21]$), a small effect (≈ 0.19 SD) that nonetheless yielded a significant one-sided directional test ($p = .033$). The Group×Dimension interaction was non-significant (Wald $\chi^2(3) = 0.79$, $p = .852$), indicating similar dehumanization profiles for women and men. One-sided, Holm-corrected tests for the individual dimensions provided no reliable evidence of gender differences (all $p_{\text{Holm}} > .44$; Table 4). Figure 4 shows that estimates for all dimensions cluster closely around zero with overlapping confidence intervals, suggesting that dehumanizing language in incel discourse is general rather than selectively targeted.

	Women	Men	Contrast	
Dimension	M (SD)	M (SD)	**Diff** (z)	**95% CI**
Moral disgust	.25 (.06)	.23 (.06)	.240	[−.143, .623]
Animal likeness	.31 (.04)	.31 (.04)	.141	[−.233, .516]
Mind denial	−.35 (.06)	−.35 (.05)	−.058	[−.437, .320]
Agency denial	−.61 (.13)	−.62 (.11)	.075	[−.117, .266]

Table 4: Women–men contrasts across dehumanization dimensions (2018–2024). Means (M) and standard deviations (SD) are computed on raw scores; contrasts report standardized differences (women > men) with 95% confidence intervals. All one-sided Holm-corrected p-values $\geq .44$.

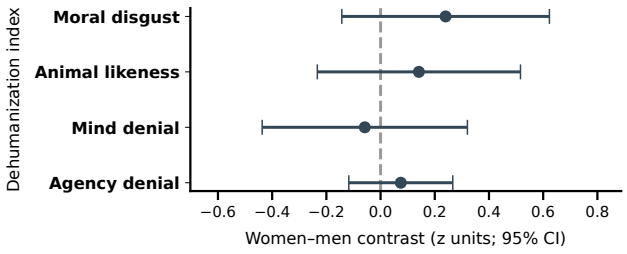

Figure 4: Standardized women–men contrasts across dehumanization dimensions (2018–24). Points show contrasts with 95% CIs; positive values indicate higher scores for women-associated labels. No intervals exclude zero.

A secondary diagnostic revealed a marked divergence in perceived social power: men-associated terms occupy significantly higher Dominance space than women-associated terms ($\beta = -0.0143$, $p < .001$; Appendix section "Diagnostic Analyses of Dehumanization Measures"). This indicates that even when dehumanizing language is present across genders, incel discourse maintains a gendered power asymmetry, positioning men closer to dominance and control.

Temporal Stability of Dehumanization Elements

To assess temporal dynamics in dehumanization (RQ3), we tested whether the women–men difference varied across years for each dimension by estimating Group×Year interactions. Table 5 reports Wald tests of these interaction terms, where χ^2 is the test statistic, df is the number of degrees of freedom, and p is the probability of observing a result at least this large under the null hypothesis of no differential change over time. Across all four dimensions, the Group×Year interaction was non-significant ($p > .05$), indicating no reliable evidence that the relative positioning of women- and men-associated terms changed significantly from 2018 to 2024. Overall, these results indicate that women–men differences in dehumanization remained stable across the study period. This pattern is also visible in Appendix section "Diagnostic Analyses of Dehumanization Measures" (Figure 9), where temporal trajectories for the dimensions of dehumanization remain broadly parallel over time.

Dimension	χ^2	df	p
Moral disgust	7.53	7	.274
Animal likeness	12.09	7	.060
Mind denial	8.22	7	.222
Agentivity index	8.59	6	.198

Table 5: Wald test results for Group × Year interactions by dehumanization dimension. Non-significant p-values indicate no evidence that women–men differences on dimensions of dehumanization changed over time (2018–2024).

Discussion

This study provides the first theory-driven, multidimensional analysis of gendered dehumanization in incel discourse. Using a seven-year corpus of 10.3 million posts and a computational framework distinguishing negative evaluation from specific elements of dehumanization (moral disgust, animal likeness, mind denial, and agency denial), we tested whether women-associated terms are evaluated more negatively, dehumanized more intensely, or show different patterns of change relative to men-associated terms.

Across analyses, we find that gender differences in dehumanization are small and inconsistent, with only a modest aggregate-level increase for women-associated terms reaching significance. No individual dimension shows a reliable women–men difference, the Group × Dimension interaction is non-significant, and these patterns remain stable from 2018–2024. Taken together, the results suggest that incel discourse is characterized by enduring, generalized hostility toward both women and men, with only weak and inconsistent evidence for gender-differentiated dehumanization.

However, these findings should not be interpreted as evidence that women and men are represented equally in incel

discourse. Extensive qualitative and discourse-analytic research documents misogynistic practices — such as sexualized insults, misogynistic neologisms, and violent narrative framings — that disproportionately target women (Marwick and Caplan 2018; Baele, Brace, and Ging 2024; Scotto di Carlo 2023; Fowler, Green, and Palombi 2023). These forms of derogation operate at linguistic layers (e.g., neologism formation, narrative framing, threat rhetoric) that our semantic indices are not designed to capture, particularly in communities that employ cryptolects and coded language (De Kock et al. 2025). Thus, while our results show no strong gender differences in semantic dehumanization as quantified here, this does not rule out the presence of gender-asymmetric misogynistic content expressed through other discursive mechanisms.

Psychological and ethnographic research helps explain this pattern. Incel ideology is grounded in grievance, nihilism, and perceived marginalization, directed outward toward women and inward toward men. Interview and survey studies show that incels often report feeling romantically excluded, socially marginalized, and treated as "subhuman" on the basis of appearance, status, or masculinity norms (Daly and Reed 2022). These perceived experiences of dehumanization are closely linked to negative emotional states (e.g., sadness, anger, despair) and are often articulated through Black Pill ideology, which frames romantic failure as biologically determined and socially enforced. Within this worldview, grievance is not only directed toward women but is also directed inward toward the self and laterally toward other men, aligning with linguistic evidence of misogyny, misandry, and intense intra-group denigration (Bogetić et al. 2023). Community-level analyses further reveal sustained expressions of sadness and hate (Coppolillo 2025), and psychological studies document elevated depression, anxiety, loneliness, and suicidal ideation among incels relative to non-incels (Daly and Laskovtsov 2022; Costello and Thomas 2025). Large-scale computational analyses similarly show that incel communities engage extensively in self-harm–related discourse and that these patterns evolve differently from mainstream mental-health communities (Ali and Zannettou 2024). Together, this literature supports our finding that dehumanization in incel discourse not only targets women but emerges from a broader framework rooted in perceived victimhood, exclusion, and moral displacement.

This interpretation aligns with large-scale analyses of gendered hate speech in online toxic communities. Coppolillo (2025) show that misogynistic and misandric language do not vary strongly by the perpetrator's gender but function as generalized discursive practices characteristic of hostile ideological spaces. Within incel forums, this hostility is embedded in a shared "red-pill" epistemology, in which members claim to have awakened to an oppressive, gynocentric social order (Marwick and Caplan 2018). Misandry is frequently invoked as a collective grievance that reinforces group identity and the belief that broader institutions are fundamentally hostile toward men (Baele, Brace, and Coan 2021). In this context, our findings suggest that dehumanization operates less as a selectively targeted process than as a broader representational pattern in which multiple social actors (women, men, and the self) are positioned as less than fully human.

Implications for Research and Intervention Methodologically, these findings highlight the value of modeling dehumanization as multidimensional rather than as a binary label. Instance-level detection struggles to capture how elements such as moral disgust, animalization, mind denial, and agency denial co-occur or persist within ideological communities, whereas theory-informed dimensional frameworks allow clearer differentiation between general hostility and representational elements (Mendelsohn, Tsvetkov, and Jurafsky 2020). From an applied perspective, multidimensional indices offer a complementary approach to detecting hate speech and abusive language online: rather than detecting isolated slurs, they characterize the broader environment in which dehumanization is embedded at scale. Because incel discourse distributes dehumanization across targets (including women, men, and 'the self') effective interventions may need to address collective narratives of grievance, status loss, and moral exclusion, not only misogynistic expressions. Finally, while our analyses are not designed to predict individual radicalization or behavioral escalation, they align with work showing that community-level representational patterns can precede overt mobilization or violence (e.g., Baele, Brace, and Coan 2021; De Kock 2024). Theory-driven NLP approaches to dehumanization may therefore help illuminate how ideological harm is sustained linguistically over time, informing research at the intersection of language, mental health, and online extremism.

Limitations Several limitations should be noted. First, our analyses rely on anonymous forum data, where user identities and demographics cannot be verified. As in prior work (Jones 2020), this limits claims about representativeness and about whether posts reflect self-identified incels versus outsiders or performative participation. Large-scale linguistic patterns therefore complement but cannot replace qualitative methods (e.g., interviews, surveys) that examine lived experience or offline behavior. Second, discourse on incel forums may be shaped disproportionately by a small number of highly active users and by reactions to salient external events, which can amplify particular themes without reflecting stable community-wide norms (Jaki et al. 2019). Third, our framework operates at the level of group-associated terms and semantic neighborhoods rather than individual users or interactional contexts. As a result, we do not capture intra-community heterogeneity, individual trajectories, or causal pathways linking dehumanizing language to downstream behaviors. Embedding-based diagnostics also reflect stable semantic associations but cannot distinguish endorsement from quotation, sarcasm, or irony. These constraints warrant caution when interpreting dehumanization scores as direct indicators of intent or harm.

Future Directions Future work could leverage recent advances in large language models to move beyond instance-level detection toward more theory-driven, multidimensional analyses of dehumanization. Although LLMs substantially outperform feature-based baselines in identifying

dehumanizing language, likely due to their ability to encode contextual and relational cues, they still struggle to distinguish dehumanization from adjacent hate speech categories, are sensitive to prompt framing, and show uneven performance across target groups (Luceri, Boniardi, and Ferrara 2023). These limitations highlight the inadequacy of binary classification approaches. An important next step is thus to use LLMs not only as detectors but as measurement tools for operationalizing theoretically grounded dimensions of dehumanization at scale (e.g., moral disgust, animalistic framing, mind denial, agency denial). Integrating LLM-based representations with dimensional frameworks may enable more precise analyses of how dehumanization elements cooccur, are selectively emphasized across targets, and persist within ideological communities, while also offering clearer diagnostics for model bias and error (Saffari et al. 2025). Although women and men were evaluated similarly negatively overall, our analyses do not show whether that negativity was expressed in similar or different ways; future work should examine this question directly. Future work could also examine whether a small number of highly active users disproportionately influence aggregate semantic estimates.

Conclusion

This study provides the first multidimensional, theory-driven analysis of gendered dehumanization in incel discourse, examining five dimensions (negative evaluation, moral disgust, animal likeness, mind denial, and agency denial) across 10.3 million posts. Building on the multidimensional framework of Mendelsohn, Tsvetkov, and Jurafsky (2020), we model dehumanization as a structured semantic configuration and find limited evidence of gender differences. Women-associated labels show a small but significant overall increase in dehumanization; however, none of the individual dimensions demonstrate reliable women–men differences. Both women and men are embedded in a persistently hostile discursive environment marked by strong negative evaluation and enduring dehumanizing associations, with gendered nuances (higher Dominance for men). Overall, the findings indicate that dehumanization in incel discourse operates as a generalized representational baseline rather than a strongly gender-differentiated pattern, underscoring the value of theory-informed, multidimensional computational measures for characterizing how ideological harm is linguistically sustained in online communities.

Ethical Statement

This study analyzes public discourse from an openly accessible online forum (Incels.is) to investigate patterns of dehumanizing and hostile language toward gendered groups, with particular attention to representations of women and men.

This research underwent independent review by an institutional Ethics Review Board as it concerns human data. All data were analyzed in accordance with ethical standards for research on publicly available online content and platform terms of use. No personally identifiable information was included in the analysis, and all examples are presented in aggregate or anonymized form. We recognize the risks associated with amplifying or re-circulating harmful language and therefore report examples sparingly and with contextualization.

As with all work on extremist or hateful content, there is a theoretical risk that the tools or insights could be misused, for example to monitor or target specific communities. To mitigate this, we release only aggregate measures and documented code, and we frame our contributions in terms of understanding dehumanization rather than optimizing harm detection for operational deployment. Our aim is to characterize and critically examine dehumanizing discourse, not to endorse, normalize, or further weaponize it.

Acknowledgments

This project was funded by the University of Melbourne Hallmark Research Initiative on Fighting Harmful Online Communication. The authors thank Hugo Lyons Keenan and Han Perry for their meticulous work in preprocessing the data, fine-tuning the language models, and extracting key information. We are also grateful to Macken Murphy for contributing manosphere-specific terminology in his role as a domain expert; to Professor Eduard Hovy and Julia Mendelsohn for their contributions to early project discussions, and to Julia for also generously sharing her Python scripts.

References

Ali, M.; and Zannettou, S. 2024. From Isolation to Desolation: Investigating Self-Harm Discussions in Incel Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 43–56.

Assenmacher, D.; Piot, P.; Laken, K.; Jurgens, D.; and Wagner, C. 2025. Beyond the Explicit: A Bilingual Dataset for Dehumanization Detection in Social Media. *arXiv preprint arXiv:2510.18582*.

Baele, S.; Brace, L.; and Ging, D. 2024. A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence*, 36(3): 382–405.

Baele, S. J.; Brace, L.; and Coan, T. G. 2021. From "Incel" to "Saint": Analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and political violence*, 33(8): 1667–1691.

Baes, N.; Haslam, N.; and Vylomova, E. 2024. A Multidimensional Framework for Evaluating Lexical Semantic Change with Social Science Applications. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1390–1415. Bangkok, Thailand: Association for Computational Linguistics.

Bandura, A. 1999. Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review*, 3(3): 193–209.

Blake, K. R.; and Brooks, R. C. 2023. Societies should not ignore their incel problem. *Trends in cognitive sciences*, 27(2): 111–113.

Bogetić, K.; Heritage, F.; Koller, V.; and McGlashan, M. 2023. Landwhales, femoids and sub-humans: Dehumanising metaphors in incel discourse. *Metaphor and the Social World*, 13(2): 178–196.

Breazu, P.; Schirmer, M.; Hu, S.; and Katsos, N. 2024. Large Language Models and the challenge of analyzing discriminatory discourse: human-AI synergy in researching hate speech on social media. *Journal of Multicultural Discourses*, 19(3): 157–175.

Bruneau, E. G.; Jacoby, N.; Kteily, N. S.; and Saxe, R. 2018. Denying Humanity: The Distinct Neural Correlates of Blatant Dehumanization. *Journal of Experimental Psychology: General*, 147: 1078–1093.

Buckels, E. E.; and Trapnell, P. D. 2013. Disgust facilitates outgroup dehumanization. *Group Processes & Intergroup Relations*, 16(6): 771–780.

Chang, W. 2022. The monstrous-feminine in the incel imagination: Investigating the representation of women as "femoids" on/r/Braincels. *Feminist Media Studies*, 22(2): 254–270.

Cikara, M.; Eberhardt, J. L.; and Fiske, S. T. 2011. From agents to objects: Sexist attitudes and neural responses to sexualized targets. *Journal of cognitive neuroscience*, 23(3): 540–551.

Coppolillo, E. 2025. Women who hate men: a comparative analysis across extremist Reddit communities. *Scientific Reports*, 15(1): 13952.

Costello, W.; and Thomas, A. G. 2025. Seeing through the black-pill: Incels are wrong about what people think of them. *Personality and Individual Differences*, 237: 113041.

Daly, S. E.; and Laskovtsov, A. 2022. " Goodbye, my friendcels": An analysis of incel suicide posts. *Journal of Qualitative Criminal Justice & Criminology*, 11(1): 1–37.

Daly, S. E.; and Reed, S. M. 2022. "I think most of society hates us": A qualitative thematic analysis of interviews with incels. *Sex Roles*, 86(1): 14–33.

De Kock, C. 2024. Jointly modelling the evolution of community structure and language in online extremist groups. *arXiv preprint arXiv:2409.19243*.

De Kock, C. 2025. Inducing lexicons of in-group language with socio-temporal context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13281–13291.

De Kock, C.; Riabi, A.; Talat, Z.; Madhyastha, P.; Schlichtkrull, M.; and Hovy, E. 2025. IYKYK: Decoding extremist cryptolects using LLMs. *Preprint*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Engelmann, P.; Trolle, P.; and Hardmeier, C. 2024. A Dataset for the Detection of Dehumanizing Language. In Chakravarthi, B. R.; B, B.; Buitelaar, P.; Durairaj, T.; Kovács, G.; and García Cumbreras, M. Á., eds., *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, 14–20. St. Julian's, Malta: Association for Computational Linguistics.

Fowler, K.; Green, R.; and Palombi, A. 2023. "From Stacys to Foids, a Discursive Analysis of the Incel's Gendered Spectrum of Political Agency". *Deviant Behavior*, 44(12): 1775–1791.

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

Giorgi, S.; Habib, D. R. S.; Bellew, D.; Sherman, G.; and Curtis, B. 2023. A linguistic analysis of dehumanization toward substance use across three decades of news articles. *Frontiers in public health*, 11: 1275975.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Gray, H. M.; Gray, K.; and Wegner, D. M. 2007. Dimensions of mind perception. *science*, 315(5812): 619–619.

Haslam, N. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3): 252–264.

Haslam, N.; and Loughnan, S. 2014. Dehumanization and infrahumanization. *Annual review of psychology*, 65(1): 399–423.

Haslam, N.; Loughnan, S.; Kashima, Y.; and Bain, P. 2008. Attributing and denying humanness to others. *European Review of Social Psychology*, 19(1): 55–85.

Haslam, N.; and Stratemeyer, M. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11: 25–29.

Hodson, G.; and Costello, K. 2007. Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychological science*, 18(8): 691–698.

Hoeken, S.; Alacam, Ö.; Fokkens, A.; and Sommerauer, P. 2023. Methodological Insights in Detecting Subtle Semantic Shifts with Contextualized and Static Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3662–3675. Singapore: Association for Computational Linguistics.

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

Jack, A. I.; Dawson, A. J.; and Norr, M. E. 2013. Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage*, 79: 313–328.

Jaki, S.; De Smedt, T.; Gwóźdź, M.; Panchal, R.; Rossa, A.; and De Pauw, G. 2019. Online hatred of women¡? br?¿ in the Incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2): 240–268.

Jones, A. 2020. *Incels and the Manosphere: Tracking men's movements online*. Master's thesis, University of Central Florida. Unpublished master's thesis.

Joshi, D. 2024. Decoding Dehumanization: Leveraging NLP to Identify Dehumanizing Language and Its Targets. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, 87–96.

Kelman, H. C. 1973. Violence without Moral Restraint: Reflections on the Dehumanization of Victims and Victimizers. *Journal of Social Issues*, 29(4): 25–61.

Lacalle, C.; Gómez-Morales, B.; Vicent-Ibáñez, M.; and Narvaiza, S. 2024. 'Seals','bitches','vixens', and other zoomorphic insults: the animalisation of women as an expression of misogyny in the Spanish Manosphere. *Cogent Arts & Humanities*, 11(1): 2298056.

Luceri, L.; Boniardi, E.; and Ferrara, E. 2023. Leveraging Large Language Models to Detect Influence Campaigns in Social Media. arXiv:2311.07816.

Marwick, A. E.; and Caplan, R. 2018. Drinking male tears: language, the manosphere, and networked harassment. *Feminist Media Studies*, 18: 543 – 559.

Mendelsohn, J.; Tsvetkov, Y.; and Jurafsky, D. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3: 55.

Mohammad, S. M. 2025. NRC VAD Lexicon v2: Norms for valence, arousal, and dominance for over 55k English terms. *arXiv preprint arXiv:2503.23547*.

Opotow, S. 1990. Moral Exclusion and Injustice: An Introduction. *Journal of Social Issues*, 46(1): 1–20.

Osgood, C. E.; May, W. H.; and Miron, M. S. 1975. *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press.

Rashkin, H.; Singh, S.; and Choi, Y. 2016. Connotation Frames: A Data-Driven Investigation. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 311–321. Berlin, Germany: Association for Computational Linguistics.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021. The evolution of the manosphere across the web. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 196–207.

Russell, J. A. 2003. Core Affect and the Psychological Construction of Emotion. *Psychological Review*, 110(1): 145–172.

Saffari, H.; Shafiei, M.; Zhang, H.; Harris, L. T.; and Moosavi, N. S. 2025. Beyond Hate Speech: NLP's Challenges and Opportunities in Uncovering Dehumanizing Language. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 26953–26968. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Salminen, J.; Almerekhi, H.; Milenković, M.; Jung, S.-g.; An, J.; Kwak, H.; and Jansen, B. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Sap, M.; Prasettio, M. C.; Holtzman, A.; Rashkin, H.; and Choi, Y. 2017. Connotation Frames of Power and Agency in Modern Films. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2329–2334. Copenhagen, Denmark: Association for Computational Linguistics.

Schweitzer, S.; and Waytz, A. 2021. Language as a window into mind perception: How mental state language differentiates body and mind, human and nonhuman, and the self from others. *Journal of Experimental Psychology: General*, 150(8): 1642.

Scotto di Carlo, G. 2023. An analysis of self-other representations in the incelosphere: Between online misogyny and self-contempt. *Discourse & Society*, 34(1): 3–21.

Sherman, G. D.; and Haidt, J. 2011. Cuteness and disgust: The humanizing and dehumanizing effects of emotion. *Emotion Review*, 3(3): 245–251.

Steuter, E.; and Wills, D. 2010. 'The vermin have struck again': Dehumanizing the enemy in post 9/11 media representations. *Media, War & Conflict*, 3(2): 152–167.

Tipler, C.; and Ruscher, J. B. 2014. Agency's role in dehumanization: Non-human metaphors of out-groups. *Social and Personality Psychology Compass*, 8(5): 214–228.

Vulić, I.; Ponti, E. M.; Litschko, R.; Glavaš, G.; and Korhonen, A. 2020. Probing Pretrained Language Models for Lexical Semantics. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7222–7240. Online: Association for Computational Linguistics.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45: 1191–1207.

Waytz, A.; Gray, K.; Epley, N.; and Wegner, D. M. 2010. Causes and consequences of mind perception. *Trends in cognitive sciences*, 14(8): 383–388.

Zsisku, E.; Zubiaga, A.; and Dubossarsky, H. 2024. Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, 241–249. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703348.

AAAI ICWSM Paper Checklist

- 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 (our analyses and interpretation focus on the domain of interest: incel communities)
- Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

- Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes (see the Methods Section)
- Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes (Discussion)
- Did you describe the limitations of your work? Yes
- Did you discuss any potential negative societal impacts of your work? Yes (see Ethical Statement)
- Did you discuss any potential misuse of your work? Yes (see Ethical Statement)
- 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 Ethical Statement)
- Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes (see Ethical Statement)
- Did you clearly state the assumptions underlying all theoretical results? NA
- Have you provided justifications for all theoretical results? Yes
- Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes (see Discussion)
- Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes (See the poor mental health in incel discourse in the Discussion section)
- Did you address potential biases or limitations in your theoretical framework? Yes (see Limitations and Future Directions)
- Have you related your theoretical results to the existing literature in social science? Yes (see Discussion)
- Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes (see Implications for Research and Intervention section in the Discussion)

Additionally, if you are including theoretical proofs...

- Did you state the full set of assumptions of all theoretical results? NA
- Did you include complete proofs of all theoretical results? NA

Additionally, if you ran machine learning experiments...

- Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes (we provide links to the associated GitHub repository: https://github.com/lraszewski/words-against-women)
- Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes (see GitHub repository)
- Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes (see Results)
- Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes (see GitHub repository)
- Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes (see Method section)
- Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No (we do not use a supervised learning approach)

Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

- If your work uses existing assets, did you cite the creators? Yes (see relevant links in paper)
- Did you mention the license of the assets? No (we rely on the original dataset's licensing; see cited sources for details)
- Did you include any new assets in the supplemental material or as a URL? Yes (see GitHub repository)
- Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes (see Ethical Statement)
- Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes (see Ethical Statement)
- If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA (we use a recently released dataset)
- If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

- Did you include the full text of instructions given to participants and screenshots? NA
- Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes (see Ethical Statement)
- Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
- Did you discuss how data is stored, shared, and deidentified? NA

Incel Dataset Descriptives

Table 6 summarizes the distribution of posts across forum categories in the incel corpus, based on available metadata. The table provides context on corpus composition and forum activity.

Target Term Lexica

Women-Referencing Terms in Incel Discourse

Complete term list: The full lexicon, illustrated by corpus frequencies in Figure 5, comprises 60 terms used in incel forum discourse to refer to women, including general descriptors, slurs, and ideologically charged language:

Forum	Posts
Inceldom Discussion	6,313,168
The Lounge	3,348,932
<forum-not-found>	623,884
Gaming	26,258
Politics, Philosophy & Religion	12,914
Must-Read Content	12,041
Anime & Manga	14,745
Other Languages	4,540
Spanish (Español)	1,134

Table 6: Number of posts per forum (metadata).

```
['female', 'females', 'foid',
'foids', 'femoid', 'femoids',
'stacy', 'stacies', 'woman',
'women', 'girl', 'girls',
'roastie', 'roasties', 'femcel',
'femcels', 'slut', 'sluts',
'whore', 'whores', 'pussy',
'hooker', 'hookers', 'becky',
'beckies', 'bitch', 'bitches',
'np', 'nps', 'jane', 'janes', 'pj',
'pjs', 'sloot', 'sloots', 'femail',
'femails', 'w0men', 'w0man',
'jailbait', 'sister', 'sisters',
'dyke', 'dykes', 'waifu', 'waifus',
'mother', 'mothers', 'mom', 'moms',
'mum', 'mums', 'she', 'she's',
'her', 'hers', 'jenny', 'jennies',
'jill', 'jills']
```

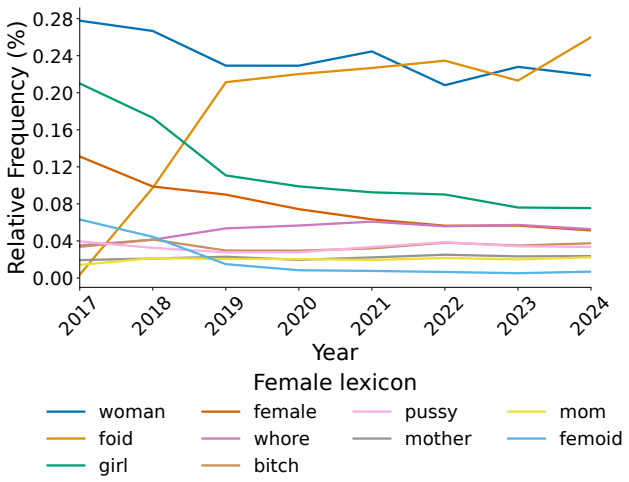

Figure 5: Annual Relative Frequency of Top Ten Terms in Women Lexicon (2017–24).

The list was initially seeded with 32 terms provided by a manosphere expert, including niche and often dehumanizing references such as "foid", "roastie", "pj", "sloot", "bangmaid", and "landwhale":

```
['foid', 'foids', 'femoid',
'femoids', 'femcel', 'femcels',
'stacy', 'stacies', 'becky',
'beckies', 'gymthot', 'jb',
'jailbait', 'landwhale',
'landwhales', 'noodlewhore',
'noodlewhores', 'pj', 'pjs', 'plain
jane', 'plain janes', 'roastie',
'roasties', 'warpig', 'warpigs',
'toilet', 'toilets', 'hole',
'holes', 'bangmaid', 'bangmaids',
'sloot']
```

This expert list was then refined inductively through corpus analysis. Low-frequency or terms whose meanings are ambiguous (e.g., "toilet", "hole", "noodlewhore") were excluded to ensure all entries were both referentially unambiguous and culturally salient. The final lexicon captures how women are categorized, sexualized, and devalued in incelosphere discourse, reflecting both linguistic function and ideological intensity.

Men-Referencing Terms in Incel Discourse

Complete term list: The following terms were used in the complete list to identify references to men in incel forum discourse, providing a basis for comparison with the women-referencing lexicon. Their corpus frequencies are illustrated in Figure 6.

```
['male', 'males', 'incel',
'incels', 'chad', 'chads',
'normie', 'normies', 'boyo',
'boyos', 'guy', 'guys', 'men',
'man', 'boy', 'boys', 'brad',
'brads', 'manlet', 'manlets',
'wizard', 'wizards', 'cuck',
'cucks', 'soyboy', 'soyboys',
'numale', 'numales', 'afc', 'afcs',
'gigachad', 'gigachads', 'jock',
'jocks', 'simp', 'simps', 'he',
'he's', 'him', 'his']
```

This lexicon spans general references to men and subculturally specific terms that carry strong evaluative connotations. These labels reflect the Incel Forums' internal hierarchies and typologies of masculinity — where men are ranked according to perceived dominance, sexual access, and conformity to group norms.

Contextual Embedding Extraction

All contextualized embeddings were extracted from the fine-tuned BERT models, as detailed in the "Contextualized Embeddings" section. We extracted contextualized embeddings using design choices that have been empirically validated for recovering *type-level lexical semantic information* from transformer language models (Vulić et al. 2020). For each occurrence of a target term, we encoded the *full sentence* containing the term using a year-specific fine-tuned `bert-base-uncased` model (i.e., a monolingual English BERT). This monolingual choice is theoretically

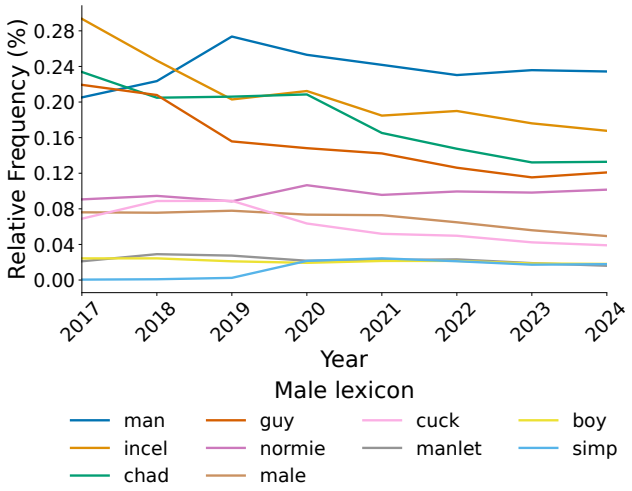

Figure 6: Annual Relative Frequency of Top Ten Terms in Men Lexicon (2017–2024).

and empirically motivated: Vulić et al. (2020) show that monolingual language models consistently retain stronger within-language lexical semantic information than multilingual models, which suffer a capacity trade-off ("curse of multilinguality") that reduces monolingual lexical quality.

Because BERT has a hard 512-token input limit, for inputs exceeding this length we extracted a window of up to 512 tokens *centered on the target occurrence*. We allocated tokens symmetrically around the target where possible (approximately 256 tokens on each side), and used an asymmetric allocation when the target occurred near a boundary (i.e., using the remaining budget on the available side). This preserves the most informative local context while respecting the model constraint.

When a target term was segmented into multiple WordPiece subword tokens, we averaged embeddings across the corresponding subwords to obtain a single token-level representation. Notably, we *excluded* special control tokens (`[CLS]`, `[SEP]`), which encode sentence-level and structural information rather than lexical meaning, from all computations, following systematic evidence from Vulić et al. (2020) that including special tokens *degrades* extracted type-level lexical semantic representations across languages, models, and layer-averaging strategies.

To recover lexical (type-level) semantics from BERT, we derived token representations by *layer-wise averaging over the lower Transformer layers* (Layers 1–6). Probing evidence from Vulić et al. (2020) indicates that (i) type-level lexical information is concentrated in lower layers, (ii) lexical knowledge is *distributed* across multiple lower layers rather than isolated in a single "best" layer, and therefore (iii) averaging across lower layers consistently outperforms single-layer extraction and averaging across all layers, which can dilute lexical information.

Finally, to obtain year-specific *type-level* embeddings, we applied an *average-over-contexts* (AOC) strategy: within each year, we averaged the contextualized token embeddings for all occurrences of a word to produce a single annual vector. This choice is directly supported by Vulić et al. (2020), who find that providing natural sentential context and averaging over occurrences (AOC) yields higher-quality type-level lexical representations than encoding words in isolation (ISO). We applied the identical extraction pipeline to target terms and to all concept-word lists (e.g., agency, purity, animal, mind), ensuring that all vectors are embedded in the same corpus- and year-specific representation space and are therefore directly comparable.

Analysis Pipeline Overview

Figure 7 summarizes the yearly analysis pipeline used in the study. Starting from the Incels.is corpus, posts were partitioned into yearly subsets (2018–2024), and a separate `bert-base-uncased` model was fine-tuned for each year. These yearly models supported three families of analyses: context-window valence estimates, dependency-based Connotation Frames analyses, and embedding-based centroid and nearest-neighbor analyses. The resulting yearly term-level estimates were then aggregated into and group–year observations for the statistical models reported in the main text.

Warriner Norms Lexicon: Valence

To assess sentiment toward target groups using the valence index introduced by Baes, Haslam, and Vylomova (2024) (see Section), we applied valence ratings from Warriner et al.'s (2013) affective norms for 13,915 English lemmas. These ratings were collected from 1,827 native English speakers (aged 16–87, 60% female), who evaluated each word on a 1–9 scale, where 1 indicated extremely negative affect (e.g., unhappy, despaired) and 9 indicated extremely positive affect (e.g., happy, contented). As Warriner et al. explain, a word's valence reflects the degree to which it evokes pleasant or unpleasant feelings in the rater.

Valence: *"You are invited to take part in the study that [...] concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. [...] The scale ranges from 1 (happy) to 9 (unhappy). At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. When you feel completely happy you should indicate this by choosing rating 1. The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored. You can indicate feeling completely unhappy by selecting 9. The numbers also allow you to describe intermediate feelings of pleasure, by selecting any of the other feelings. If you feel completely neutral, neither happy nor sad, select the middle of the scale (rating 5)."*

Connotation Frames Lexica

Perspective lexicon (Rashkin, Singh, and Choi 2016): We use the public Connotation Frames resource (around 940 English verbs), taking only `Perspective(ws)` (writer→subject) and `Perspective(wo)` (writer→object). Scores range from $[-1, 1]$ where higher scores reflect a more positive

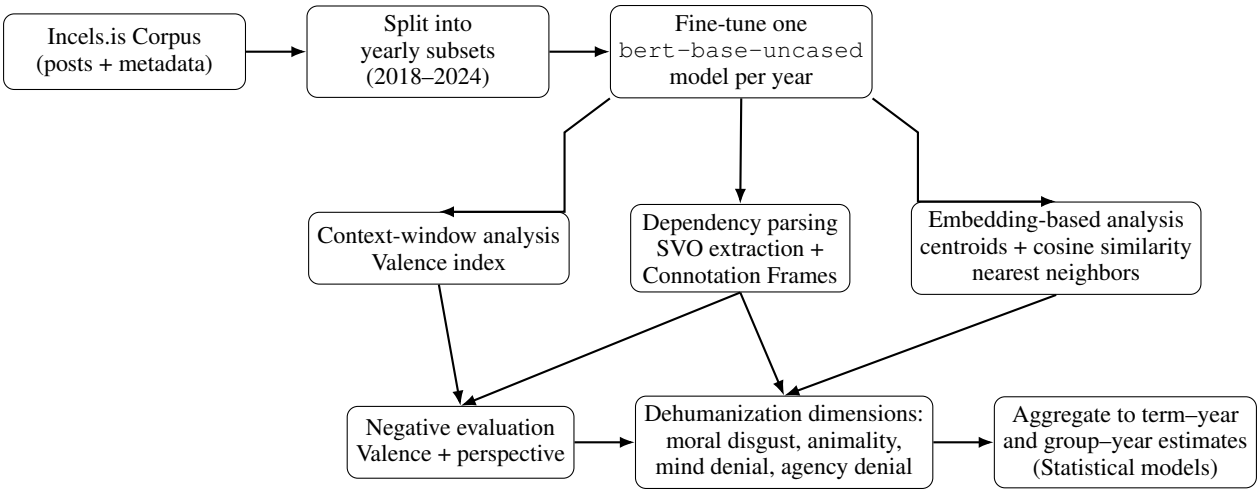

Figure 7: Schematic overview of the yearly analysis pipeline. The corpus was partitioned into yearly subsets, a separate `bert-base-uncased` model was fine-tuned for each year, and each yearly model was used to compute context-based, dependency-based, and embedding-based measures. These were aggregated into term–year and group–year observations for the statistical analyses reported in the main text.

writer stance. Annotations were crowd sourced on Amazon Mechanical Turk with 15 judgments per verb; categorical labels were mapped to {+1.0, +0.5, 0.0, -0.5, -1.0} and averaged. Practical bands: $[-1, -0.25]$ negative, $[-0.25, 0.25]$ neutral/mixed, $[0.25, 1]$ positive. For illustration, the verb *make* has `Perspective(ws)`$\approx$ 0.10 and `Perspective(wo)`$\approx$ -0.07: if the target appears as the subject in an SVO (we record $+0.10$), but if it appears as the object we record -0.07. Likewise, *give* has `Perspective(ws)`$\approx$ 0.40 and `Perspective(wo)`$\approx$ 0.23, yielding a larger positive score when the target is the subject than when it is the object.

Manual inspection of high- and low-scoring Connotation Frames tuples confirmed that extreme scores correspond to intuitively negative versus positive writer stance toward group referents. Highly negative scores are typically associated with constructions in which women- or men-associated labels occur as the grammatical subject of negatively valenced predicates (e.g., *women commit sui [i.e., suicide]*, *he committing evil incels*), while highly positive scores arise in agentive or prosocial frames in which targets are depicted as rescuers, saviors, or speakers (e.g., *my mother saved her*, *two men rescue his assailant*). The best role was often the subject. These patterns indicate that CF perspective scores capture directed evaluative stance grounded in predicate semantics rather than surface sentiment alone.

Agency and power lexicon (Sap et al. 2017). We use the Connotation Frames of Agency and Power lexicon (Sap et al. 2017), which assigns each verb (listed in 3rd-person singular form) an *agency* label {`pos`, `neg`, `equal`} and a *power* label {`agent`, `theme`, `equal`}. The resource covers nearly 2,000 transitive and intransitive verbs and was annotated via Amazon Mechanical Turk with 3 judgments per item using placeholder templates (e.g., "X alters Y"); majority/consensus labels were released. In our analysis, we use the agency facet to characterize the agency attributed to subjects in SVO tuples (power labels are retained for reference).

NRC-VAD Lexica: Valence and Dominance

For analyses requiring affective scores linked to token-level embeddings, we used Valence (see Section) and Dominance (see Section) ratings from the NRC VAD Lexicon v1.0 (Mohammad 2025), which provides crowd-sourced scores for approximately 20,000 English words. Valence captures affective positivity or negativity, while dominance reflects perceived control, competence, and power. The lexicon was constructed from terms drawn from the union of several established English lexical resources, including:

- **NRC Emotion Lexicon**: 14,000 words annotated for eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), based on high-frequency content words from the Google N-gram Corpus.
- **General Inquirer**: all 4,206 terms from its positive and negative affect lists.
- **ANEW**: all 1,061 terms from the Affective Norms for English Words.
- **Warriner et al. Lexicon**: all 13,915 terms rated for valence, arousal, and dominance.
- **Roget's Thesaurus Categories**: 520 words corresponding to Plutchik's eight basic emotion categories.
- **Hashtag Emotion Corpus (HEC)**: 1,000 high-frequency words and emoticons from tweets tagged with emotion-related hashtags (e.g., #anger, #sadness).

These combined sources served as the basis for VAD annotation in the NRC VAD Lexicon. The present study only uses Valence and Dominance scores, outlined below.

Valence *"Consider positive feelings (or positive sentiment) to be a broad category that includes: positiveness / pleasure / goodness / happiness / greatness / brilliance / superiority / health etc. Consider negative feelings (or negative sentiment) to be a category that includes: negativeness / displeasure /badness / unhappiness / insignificance / terribleness / inferiority / sickness etc. If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the Merriam Webster) or on the internet. Quality Control Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. An occasional misanswer is okay. However, if the rate of misanswering is high (e.g., >20%), then all of one's HITs may be rejected. Select the options that most English speakers will agree with.*
Q1. <term> is often associated with:
3: very positive feelings
2: moderately positive feelings
1: slightly positive feelings
0: not associated with positive or negative feelings
-1: slightly negative feelings
-2: moderately negative feelings
-3: very negative feelings "

Dominance *"This task is about words and their association with dominance, competence, control of situation, or powerfulness. Consider dominance, competence, control of situation, or powerfulness to be a broad category that includes: dominant, competent, in control of the situation, powerful, influential, important, autonomous, etc. Consider submissiveness, incompetence, controlled by outside factors, or weakness to be a broad category that includes: submissive, incompetent, not in control of the situation, weak, influenced, cared-for, guided, etc. This task is not about sentiment. (For example, something can be positive and weak (such as a flower petal) and something can be negative and strong (such as tyrant)."*

Nearest-Neighbor Extraction

We use nearest-neighbor (NN) extraction as a shared embedding-based procedure supporting both measurement and diagnostics. All NN computations operate on the year-specific type embeddings.

Vocabulary For each year, we constructed a candidate vocabulary by ranking word types by frequency, removing NLTK stop words and additional low-information items, and retaining up to the 8,000 most frequent remaining tokens. This yields a content-bearing vocabulary that is both semantically diverse and computationally tractable.

Group Centroid Construction Women and men centroids were computed per year by averaging the type embeddings of all group labels in the lexicon. All similarity measures are within-year and based on a fine-tuned model.

Nearest Neighbors For each centroid, we ranked all candidate vocabulary items by cosine similarity and retained the top 500 neighbors, excluding the centroid's own labels. These NN sets characterize the local semantic neighborhoods of women- and men-associated terms.

Downstream Use NN sets serve two roles: (1) Qualitative/diagnostic: we summarize raw neighbor lists to inspect local semantic structure (e.g., Appendix I); (2) Lexicon-restricted aggregation: we intersect NN sets with external lexicons (e.g., NRC-VAD) and compute cosine-weighted averages of valence, arousal, or dominance. For theoretically motivated dimensions based on concept dictionaries (moral disgust, animal likeness, mind denial), we compute centroid–centroid similarity directly.

Negative Evaluation of Target Groups: Robustness and Diagnostic Analyses

Figure 8 provides robustness checks using NRC-VAD valence norms. Panel (A) recomputes the neighboring-lemma valence index with NRC-VAD and replicates the downward trend from the main text, confirming that increasing negativity is not driven by lexicon choice. Panel (B) shows a nearest-neighbor centroid analysis averaging NRC-VAD valence across the 500 closest neighbors to each group centroid; unlike the lemma- and frame-based indices, this centroid score trends upward over time. As nearest-neighbor methods capture broad semantic drift not directed evaluative meaning, this measure reflects global semantic association.

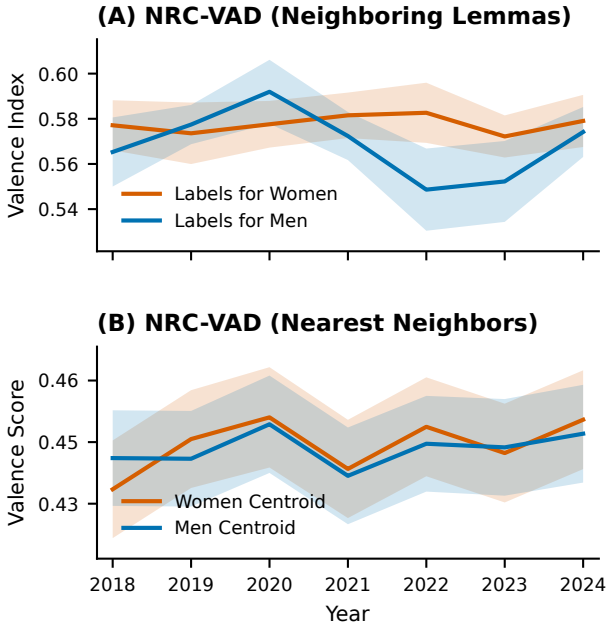

Figure 8: Robustness and diagnostic analyses of valence using NRC-VAD valence norms. Panel (A) recomputes the contextual valence index, showing mean valence of neighboring lemmas occurring within ±5 tokens of each group label, averaged across terms. Panel (B) reports a nearest-neighbor centroid analysis, in which the 500 closest neighbors to each group centroid are averaged for valence.

Note. Panel (A) reproduces the main downward trend, supporting robustness to lexicon choice. Panel (B) reflects broader semantic neighborhood and shows upward trend. Ribbons indicate ±1 SEM.

Nearest Neighbors to Target Groups

Because group centroids are constructed by averaging embeddings of multiple group-related terms, nearest neighbors reflect the broader semantic neighborhood surrounding each centroid. In this analysis, all target terms used to construct the centroids are explicitly excluded from the neighbor search, ensuring that the tables summarize associative and affective correlates (e.g., evaluative adjectives, mental state descriptors, and behavioral predicates) rather than trivial morphological variants. Nearest neighbor tables (Tables 7 and 8) therefore capture local semantic structure in the embedding space at each time point and are interpreted as descriptive diagnostics of how group concepts are situated in discourse, rather than as evidence of centroid displacement or sense change over time.

Women Centroid

2017	2018	2019	2020	2021	2022	2023	2024
dislikes	dislikes	dislikes	dislikes	dislikes	dislikes	snooty	snooty
sneering	snooty	snooty	lusty	snooty	snooty	dislikes	dislikes
snooty	planky	sneering	sneering	lusty	lusty	lusty	concerningly
sado	sneering	obsess	imagines	sneering	drowsy	sneering	lusty
obsess	goalie	goalie	lovelessness	enjoyer	sneering	drowsy	drowsy
plankton	mentalist	devotee	snooty	obsess	obsess	enjoyer	enjoyer
self-esteem	lusty	imagines	drowsy	abuser	imagines	imagines	intellectuality
drowsy	abuser	abuser	blushy	drowsy	enjoyer	abuser	antsy
disliking	imagines	abusiveness	catty	self-esteem	devoteeism	obsess	imagines
brainwashing	obsess	drowsy	goalie	dazzler	abuser	self-esteem	sneering

Table 7: Top ten nearest neighbors for the *women* group centroid in each year (2017–24), excluding the target terms.

Men Centroid

2017	2018	2019	2020	2021	2022	2023	2024
sado	planky	snooty	dislikes	snooty	snooty	snooty	snooty
snooty	snooty	dislikes	lusty	dislikes	lusty	brainy	brainy
dislikes	dislikes	sneering	sado	enjoyer	enjoyer	aspirer	antsy
plankton	sneering	antsy	snooty	planter	planter	enjoyer	enjoyer
sneering	antsy	imagines	imagines	brainy	dislikes	planter	dislikes
obsess	goalie	obsess	sneering	lusty	drowsy	antsy	concerningly
imagines	brainy	brainy	dazzler	malevolent	dazzler	dislikes	aspirer
abuser	imagines	abuser	goalie	dazzler	sneering	lusty	lusty
brainy	malevolent	thinker	brainwasher	abuser	malevolent	brainer	careerist
embarrass	brainer	admirer	antsy	sadomasochism	imagines	feeler	intellectuality

Table 8: Top ten nearest neighbors for the *men* group centroid in each year (2017–24), excluding the target terms.

Diagnostic Analyses of Dehumanization Measures

Aggregate Trends

Figure 9 provides a consolidated visual summary of dehumanization-related semantic trends across all measured dimensions.

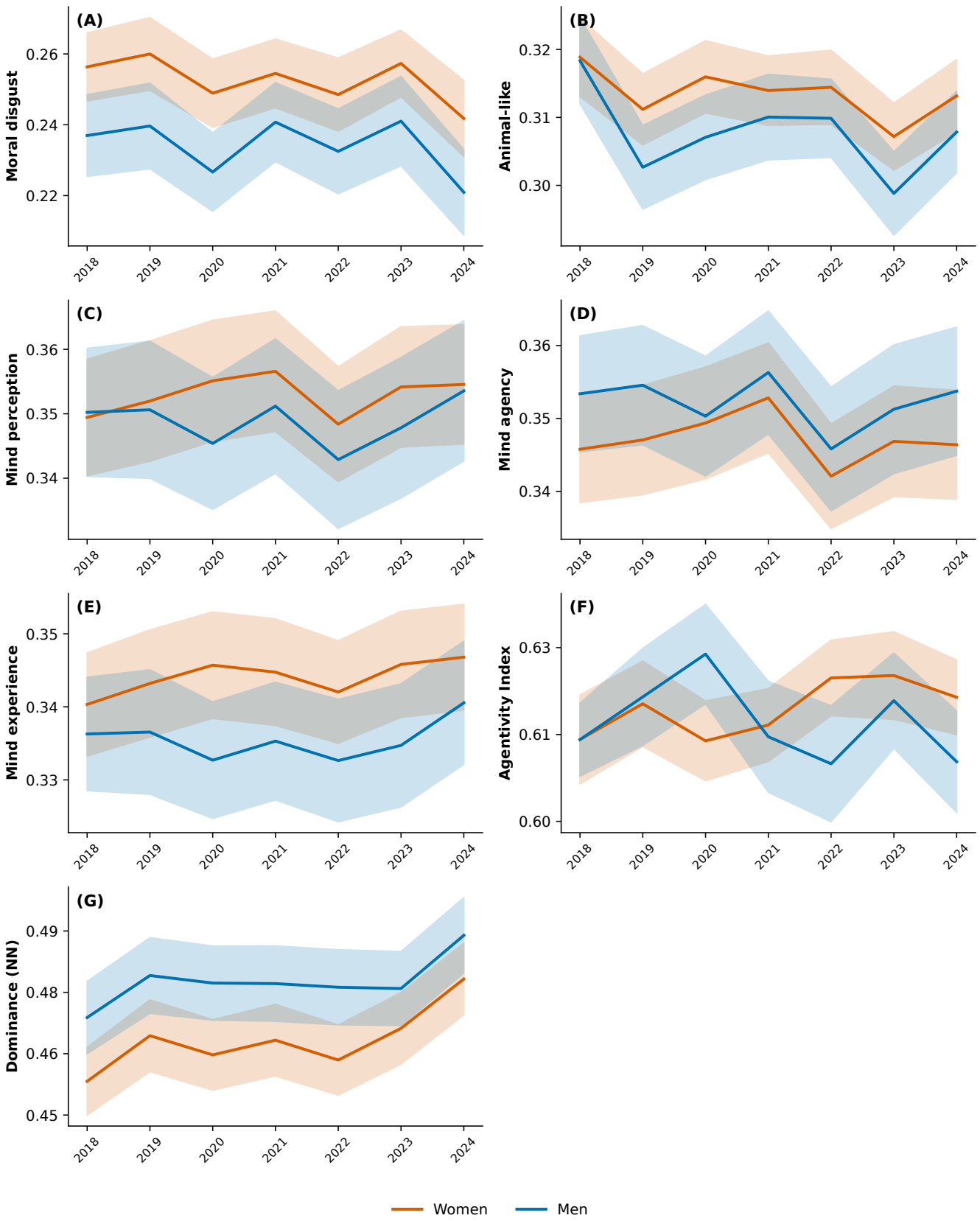

Figure 9: Dehumanization-related semantic trends in incel discourse (2018–24). Panels show yearly mean scores.