IDA - Incel Data Archive: a multimodal comparable corpus for exploring extremist dynamics in online interaction

Selenia Anastasi, Tim Fischer, Florian Schneider, Chris Biemann
University of Genoa, Language Technology Group (Hamburg University)
selenia.anastasi@edu.unige.it, tim.fischer@uni-hamburg.de,
florian.schneider-1@uni-hamburg.de, biemann@informatik.uni-hamburg.de

Abstract

Extremist online communities are rapidly growing locally, posing potential threats to European and non-European countries. To gain insight into the dynamics of interaction within these web-based extremist groups, we present IDA, the Incel Data Archive. IDA is a multilingual and multimodal corpus compiled from Incel forums in both Italian and English languages. With its collection of forums, blogs, and websites, the Incelosphere serves as an ideal case study for examining interaction dynamics within extremist online communities from a cross-cultural perspective. Therefore, our work makes a twofold contribution: firstly, it provides an original cross-cultural perspective on the Incel phenomenon, and secondly, it extensively discusses the challenges and opportunities encountered when constructing a multimodal and multilingual corpus from discussion forums. To achieve this, we employ a mixed-method approach to Computer Mediated Communication. In order to shed light on important differences between the two communities, we conducted an exploratory analysis based on a novel topic modeling technique based on Transformer architectures. This approach allowed us to delve into the themes present in the two corpora. The results of our thematic exploration demonstrate not only variations in the discussion topic favoured by each community but also differences in the targets of their hateful content.

Keywords: CMC corpora, Incels, Online Extremism, Multimodality, Multilingual Corpora

1. Introduction

After spreading within Reddit, Incels communities gradually aggregated outside mainstream social networks, creating the formation of an independent insular cluster of local-based communities. Recently, several studies (Gillett and Suzor, 2022; Trujillo and Cresci, 2022) supported the hypothesis that moderation and quarantine practices adopted by mainstream platforms, may foster the growth of hateful insular peripheral communities akin to echo chambers. The creation of the dataset presented in this work was motivated by the need to draw upon spontaneous examples of Computer Mediated Discourse that exhibited similar content from various perspectives, framing this phenomenon at a local level. Moreover, even though the discourse of the Incelosphere is characterised by its hateful, misogynistic and anti-feminist contents, we argue that a corpus consisting of data from the Incelosphere may be useful in answering broader research questions that address the general understanding of the digital ecosystem in which extremist users interact. Our contribution is thus twofold: first, we intend to contribute to a deepen understanding of Incel communities from a cross-cultural perspective. Secondly, as datasets from these sources have not yet been made openly available for academic purposes, this study aims to fill this gap by addressing some of the challenges that accompany the construction of a multimodal and bilingual corpus in Italian and English. From a methodological perspective, we intend to offer our perspective and solutions to aid in the construction of a corpus intended to study of the forums-based communities by drawing on the Computer Mediated Discourse research field. We believe that this perspective is particularly relevant because it considers both the level of user interaction, the *affordances* of the discussion fora, as well as how the community and its sociocultural context influence each other.

2. The Incelosphere so far

Anglophone Incel communities have been studied from a wide variety of perspectives, ranging from psychology to discourse analysis. Many of these studies were focused on Reddit groups (r/ForeverAlone and r/Incels subreddits), which are archived in datasets and can be used as corpora. In Sociology, studies focused on the discursive practices, rhetoric and argumentation style, symbolism, and sexual imagery of Incel communities(Massanari, 2017; Waśniewska, 2020; Tranchese and Sugiura, 2021; Aiston, 2023; Prażmo, 2022), male and female identity construction (Ging, 2019; Chang, 2022; Thorburn et al., 2023), target of the hateful content (Pelzer et al., 2021), thematic and rhetorical connections to far-right oriented groups (Nagle, 2017), anti-feminism, values, normative orders, and group beliefs(Sugiura, 2021; O Malley et al., 2022; Heritage and Koller, 2020). Empirical analyses and terrorism studies have sought to trace, also through dynamic cross-platform approaches, the development of violent extremism in the main Anglophone communities (Ribeiro et al., 2021; Baele et al., 2023), as well as their misogynistic stances (Jaki et al., 2019; Farrell et al., 2019). For Baele and colleagues, "incel discourse demonstrates typical markers of extremist language" that is "an essentialist categorisation of society into sharply delineated *in-groups* and *out-groups* where the latter are linguistically dehumanised, and a conspiratorial narrative presenting the in-group as the victim of an all-powerful structure of oppression" (Baele et al., 2023). Moving from the latter consideration, our study aims to frame the studied communities as insular clusters that have spontaneously arisen among individuals who, while sharing the same ideology and similar ways of articulating it, may display different levels of extremism.

3. Online discussion forums and Computer Mediated Discourse

The Computer Mediated Discourse (CMD) approach traditionally concerns the study of discourse in interactions where communication occurs through computers or mobile devices (Herring and Androutsopoulos, 2015). While much of the research has focused on texts, recent attempts have been made to incorporate graphic, audio, and video elements, as well as stylistic and stylometric elements at the level below the utterance. Additionally, the CMD approach distinguishes itself from other discourse approaches by considering the importance of platform-specific affordances and how they shape interaction, an aspect we aimed to preserve in our corpus. Indeed, forums-mediated conversations are not simply digitised conversations, but rather a distinct type of interaction with their own conditions of production and interpretation. Nonsynchronous digital interaction promotes the presence of complex sequential organisations, with connections to previous shifts and the management of multiple lines of interaction in parallel. This necessitates participants to develop new methods for indexing sequential connections, self-introduction, greetings, and attention calls. Taking these aspects into account, the primary objective of this study is to illuminate the language and dynamics of interaction within Incel extremist communities, bridging the gap in resources that are openly available and can be used to examine this phenomenon from a cross-cultural perspective. The specificity of the corpus should not come as a surprise. With the spreading of new social networks sites such as TiKTok, and the growing interest in particular phenomena related to digital communication, we have witnessed the development of several corpora tailored for specific purposes in recent years. As generic linguistic corpora such as the WaCky corpus(Baroni et al., 2009) do not enable researchers to delve into specific topics, more recent studies have focused on creating corpora from online content related to specific themes, such as anti-vaccine movements, fake news, and conspiracy theories (Miani et al., 2021). In the next paragraph, we offer a more detailed description of the corpus design, data collection criteria and annotation processes.

4. Corpus constructions

4.1. Collection criteria

To ensure that the samples between the two language-based macro communities are homogeneous, both in terms of characteristics of the medium and local situational factors, we took as a starting point the affordances offered by all the forums examined and on the similarities between the discussion topics, and the user's identity claim as Incel. The selection of the forum was carried out with qualitative methods, including expert-domain close reading, for the purpose of analyzing the similarities between the two communities and identity claims their user-base. Thus, we selected only those forums that showed the greatest similarity in structure, affordances and purposes. According to relevant literature (Lilly, 2016), we considered the different communities present within the Manosphere (PUA, MGTOW, MRA, etc.), and assessed the different user-reported positioning and framing with respect to issues of masculinity and anti-feminism. Having to place each of these groups on an ideal continuum ranging from "not at all toxic" to "very toxic", according to (Farrell et al., 2019) and (Ribeiro et al., 2021), Incels Anglophone communities shows a sharp rise in the mean toxicity score compared to PUAs and MRAs. For this reason, we believe that researching the Incel forums may be a worthy case study for a cross-cultural investigation on the rise of new online extremism. After the selection of the forums, we define relevant sections and threads according to our purposes. We chose to select and collect only specific freely accessible threads that did not require any formal subscription to the two forums. This was due to two main reasons: first, the ethical one - avoiding to violate the privacy policies of the platform; second, to reduce the risk for the researchers to be subjected to potential violence and other forms of retribution.

4.2. Dataset collection

We collected the data and processed the dataset using well-established methods (Holtz et al., 2012). Both forums are structured hierarchically in sections, threads, and posts. Every section can contain a varied number of threads of different lengths that relate to roughly one topic, and consisting of asynchronous conversation flows in which can involve various users.

For the composition and collection of the dataset, we implemented multiple crawlers, one for each forum, in order to systematically download threads and posts of the sections of our interest. Given the URLs to the sections of interest (e.g. Introduction, Inceldom Discussion, Off-Topic), the crawler performs the following steps:

1. Visit each section. Collect URLs to all threads of that section
2. Visit every thread. Extract metadata of the thread and collect URLs to all of its posts.
3. Visit every post. Extract metadata of the post and its content. If available, download linked materials such as image, video or audio data.

To be more specific, the crawlers extract `title`, `permalink`, `date` and `id` for threads, and `speaker`, `content`, `permalink`, `date`, `id`, `thread id`, `title`, `image urls` and `reply to` for posts. With this procedure, the created dataset captures the hierarchical structure of the forums of sections, threads and posts as well as the conversational flow of the threads and posts of referring, citing and replying to other users. The final dataset comes in various formats, including CSV, JSON, HTML and PDF. The two latter formats are a reconstruction of the original forum format and well suited for annotation tasks. The crawlers are implemented with Scrapy[1], a Python framework for extracting data from websites. To navigate the forums and to extract metadata, content or linked materials, it is required to specify CSS and XPath selectors that point directly to the desired content. These identifiers are

[1] https://scrapy.org/

specific to every website and forum, which makes the development of such crawlers a careful and time-consuming endeavour.
The English dataset is about 10 times the size of the Italian. Further statistics of the crawled datasets can be seen in Table 1.
A phase of post-processing has been devoted to managing external links and videos embedded in user posts. These information have been automatically replaced by appropriate labels. A second challenge involved the anonymization of user names in threads and posts to ensure data privacy.

	English	Italian
Number of threads	369.174	35.624
Number of posts	7.359.727	740.278
Average posts per thread	20	21
Average post lengths (in chars)	161,45	281,90
Number of images (total)	425.259	20.183
Number of images (unique)	72.22%	93.69%

Table 1: Main statistics of the English and Italian datasets.

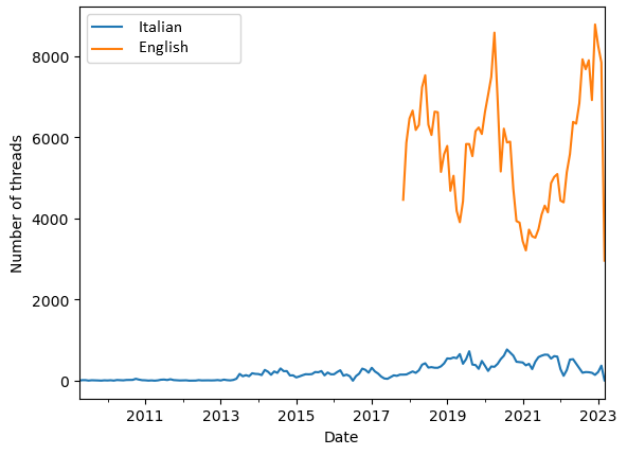

Figure 1: Number of new threads over time for English and Italian forum.

5. Corpus exploration

5.1. Methodology

The preliminary exploration and comparison of the dataset contents was exectured in two phases. The first phase involved the use of topic modeling for extracting topics through an unsupervised approach. The second phase employed a Corpus Assisted Discourse Analysis, following the steps outlined in Baker's proposed model (Baker, 2006): Description, Interpretation, Explanation, and Evaluation. The Explanation phase incorporated cross-referencing with the data from other sources such as newspapers and findings from previous research, particularly concerning the Anglophone community. Since the Italian community is studied to a lesser extent, the collection of all the relevant information has been carried out during the course of the last year by periodically accessing the forum and investigating the practices of the community in close-reading. This qualitative analysis focused on a thorough reading and revolved around identifying key actors and topics being discussed within the forum, with a specific emphasis on aspects potentially influenced by sociodemographic variables such as entertainment, sexuality, and employment. The interpretation phase for the topics that were generated by the topic modeling was supported by exploring the meanings of the keywords contained within each topic list using the concordance tool in Sketchengine, on a subcorpus of both datasets consisting of approximately 3 million words each. Alongside the description of the dataset, our work, we showcases the potential for future research based on the Incelosphere corpus.
For the first phase, regarding the topic modeling, we randomly sampled 10% of the threads from the English forum (36.917), balancing the smallest Italian corpus (35.624). Although criticised (Brookes and McEnery, 2019), in Social Sciences and Digital Humanities, a widely used technique for exploring large unlabeled corpora is topic modeling. In our analysis, we replaced the classic approach based on bag-of-words representations and LDA (Blei et al., 2003) with a new approach based on transformer architectures (Vaswani et al., 2017), which allows for the extraction of words not only in relation to their distribution throughout the documents, but also in relation to their context of occurrence. Topic modeling based on BERT embeddings (Grootendorst, 2022) proved to be reliable for its high versatility and stability across domains, the possibility to perform analysis on multilingual data, and the ability to automatically extract the appropriate number of topics based on the sample size (Egger and Yu, 2022). This allowed us to obtain highly disambiguated word lists and minimised output manipulation. We used the Sentence Transformer (Reimers and Gurevych, 2019) model ALL-MPNET-BASE-V2[2] to compute vector representations of the threads, as it yielded the best clustering in our experiments. Topic modeling allowed us to obtain some preliminary insights on the topic trends, both synchronically and diachronically (see Fig.2).

5.2. Cross-media and cross-cultural analysis

The first step in analysing the contents of both datasets has been to visualise the flow of new messages over time (see Fig. 1). We found that the flow of new threads differs significantly between the two forums. The Italian forum displays a relatively stable pattern of new messages per day, whereas the English forum exhibits distinct peaks in 2018, 2020, and 2023, as well as a notable decrease in 2019 and 2021. The reason behind this trend remains to be accurately determined; however, from a cross-media perspective, we notice a correspondence between the decrease of messages and how the media gives attention to this community cyclically, mostly when there are crime events associated to it. Notably, there have been 50 documented cases of incel violence since 2014, including the murder of five people by Jake Davison in Plymouth in August 2021 and Gabrielle Friel's weapons stockpiling in 2019 for a terror-

[2]https://www.sbert.net/docs/pretrainedmodels.html

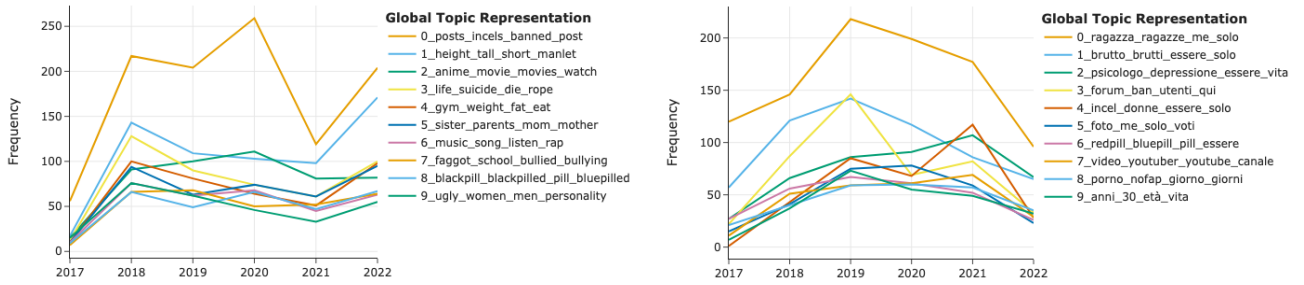

Figure 2: Dynamic topic modelling comparison of the English (Top) and Italian (Bottom) forum. Best viewed digitally with colour and zoom.

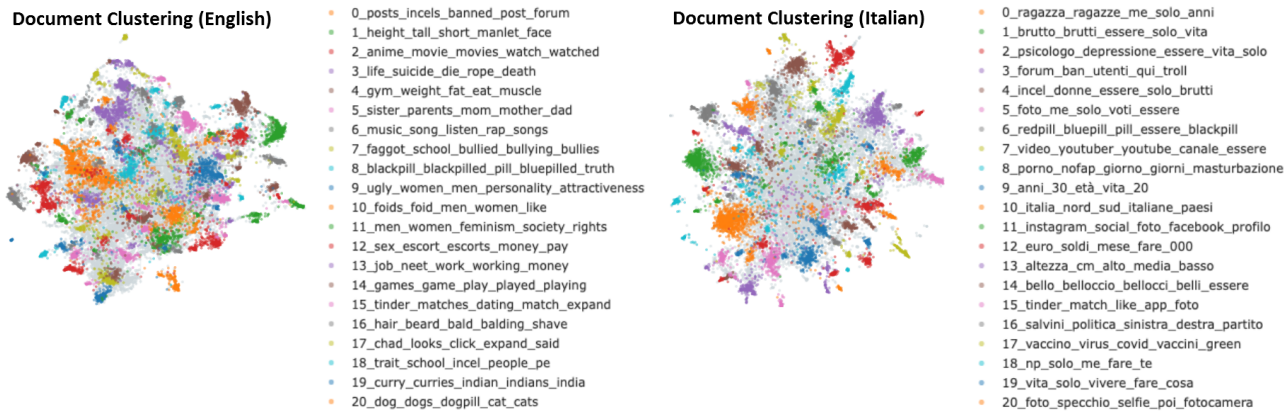

Figure 3: Static topic modelling comparison of the English (Left) and Italian (Right) forum. Best viewed digitally with colour and zoom.

ist attack in Scotland[3]. Furthermore, according to previous literature (Baele et al., 2023), we found that both, in Anglophone and Italian groups, there was an increase in the flow of messages in correspondence with the pandemic and post-pandemic years. The relationship between media attention and the growth of online extremist communities has already been observed elsewhere (Sugiura, 2021), but the correlation can be better supported by further and more in-depth analyses.

Results from the dynamic topic modeling (see Fig. 2) contain some clues about their differences: in the Anglophone community, main topics are mostly related to the user's activity (both internal and external); mainstream entertainment such as movies and music; aesthetics issues, particularly height and weight; reference to bullying, suicide, death and rapes, as well as reference to feminine family components (sisters and mothers) and women in general. Interestingly, in the Italian forum the 2017 marks a turning point in user interest. Prior to 2017, the most frequent theme appears to have been the identity traits that characterize the user base and gives the group its name (*being ugly*), while after 2017, discussions are directed towards girls and women (or in slang, "non-persons"), anti-feminism, loneliness, and mental health. Topics concerning the internal life of the forum are still present, such as "ban" and "users", indicating possible concerns on boundary maintenance.

The static clustering of the two datasets (see Fig. 3) shows the prevalence of social and affective concerns in the Anglophone group, such as unemployment, family care, sexuality and prostitution. This also emerges from the the Italian forum, where mainstream platforms such as YouTube and Instagram appear to play a prominent role. Moreover, from the Italian data in particular, aesthetic evaluation seems to be the prevailing community practice. This is not surprising, confirming the cornerstones of Incel's theories in the so-called "LMD theory", acronym for Look, Money, and Status, according to which both men and women are considered, and consider themselves, "as sexual objects to be evaluated and inserted in a hierarchical order characterised mainly by aesthetics" and economical status (Dordoni and Magaraggia, 2021). The widespread reference to ethnic categorisations such as "white", "black", "Indians" (or the incel slang variant, "curry/curries"), "Jews", and "Asians", along with keywords such as "race" and "ethnicity" in the Anglophone group is worthy of further investigation. In contrast, this pattern does not seem to emerge in the Italian forum, where stereotypes address the difference between men and women of southern and northern Italy. This aspect marks a point of continuity between the two communities, and can provide important clues for future analyses aimed at revealing mutual influence between the cultural ground of the user base and their radical instances.

6. Future Works

For further analyses, we plan to apply computational techniques such as network analyses and in-depth hate speech detection. These analyses can provide additional insights

[3]https://www.theguardian.com/lifeandstyle/2021/mar/03/incel-movement-terror-threat-canada

both on the linguistic level (is there any difference in the way hate is expressed between the two linguistic communities? Who are the target groups?) and on the level of social structures and internal hierarchies. Finally, we plan to annotate the textual data in order to reveal interactional and rhetorical patterns, while the images will be annotated to provide a new benchmark for misogyny recognition.

7. References

Aiston, J. (2023). *Argumentation strategies in an online male separatist community*. Ph.D. thesis, Lancaster University.

Baele, S., Brace, L., and Ging, D. (2023). A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence*, pages 1–24.

Baker, P. (2006). *Using corpora in discourse analysis*. Bloomsbury Publishing.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brookes, G. and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

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

Dordoni, A. and Magaraggia, S. (2021). Modelli di mascolinità nei gruppi online incel e red pill: Narrazione vittimistica di sé, deumanizzazione e violenza contro le donne. *AG About Gender-International Journal of Gender Studies*, 10(19).

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7.

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96.

Gillett, R. and Suzor, N. (2022). Incels on reddit: A study in social norms and decentralised moderation. *First Monday*.

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

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Heritage, F. and Koller, V. (2020). Incels, in-groups, and ideologies: The representation of gendered social actors in a sexuality-based online community. *Journal of Language and Sexuality*, 9(2):152–178.

Herring, S. C. and Androutsopoulos, J. (2015). Computer-mediated discourse 2.0. *The handbook of discourse analysis*, pages 127–151.

Holtz, P., Kronberger, N., and Wagner, W. (2012). Analyzing internet forums. *Journal of Media Psychology*.

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of womenin the incels.me forum. *Journal of Language Aggression and Conflict*, 7(2):240–268.

Lilly, M. (2016). *'The World is Not a Safe Place for Men': The Representational Politics Of The Manosphere*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Massanari, A. (2017). # gamergate and the fappening: How redditâs algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346.

Miani, A., Hills, T., and Bangerter, A. (2021). Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.

Nagle, A. (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

O Malley, R. L., Holt, K., and Holt, T. J. (2022). An exploration of the involuntary celibate (incel) subculture online. *Journal of interpersonal violence*, 37(7-8):NP4981–NP5008.

Pelzer, B., Kaati, L., Cohen, K., and Fernquist, J. (2021). Toxic language in online incel communities. *SN Social Sciences*, 1:1–22.

Prażmo, E. (2022). In dialogue with non-humans or how women are silenced in incelsâ discourse. *Language and Dialogue*, 12(3):383–406.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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, pages 196–207.

Sugiura, L. (2021). *The incel rebellion: The rise of the manosphere and the virtual war against women*. Emerald Group Publishing.

Thorburn, J., Powell, A., and Chambers, P. (2023). A world alone: Masculinities, humiliation and aggrieved entitlement on an incel forum. *The British Journal of Criminology*, 63(1):238–254.

Tranchese, A. and Sugiura, L. (2021). âi donât hate all women, just those stuck-up bitchesâ: How incels and mainstream pornography speak the same extreme language of misogyny. *Violence against women*, 27(14):2709–2734.

Trujillo, A. and Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Waśniewska, M. (2020). The red pill, unicorns and white

knights: Cultural symbolism and conceptual metaphor in the slang of online incel communities. *Cultural conceptualizations in language and communication*, pages 65–82.