Exploring digitally-mediated communication with corpora

Digital Linguistics

Edited by
Andreas Witt

Volume 2

Exploring digitally-mediated communication with corpora

Methods, analyses, and corpus construction

Edited by
Louis Cotgrove, Laura Herzberg and Harald Lüngen

DE GRUYTER

The Open Access version of this publication was funded by the Leibniz Association's Open Access Publishing Fund for the promotion of scientific research.

ISBN 978-3-11-143259-5
e-ISBN (PDF) 978-3-11-143401-8
e-ISBN (EPUB) 978-3-11-143433-9
ISSN 2751-1278
DOI https://doi.org/10.1515/9783111434018

Library of Congress Control Number: 2025936015

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Cover image: piranka/E+/Getty Images
Printing and binding: CPI books GmbH, Leck

www.degruyterbrill.com
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com

Table of contents

Selenia Anastasi, Tim Fischer, Florian Schneider, and Chris Biemann

IDA – Incel Data Archive

A multimodal comparable corpus for exploring extremist dynamics in online interaction

Abstract: Extremist online communities connected to the male supremacist ecosystem are rapidly growing worldwide in insular groups outside the mainstream social network sites, which is a concerning phenomenon on a global scale. To understand the dynamics of these groups, we introduce IDA, the Incel Data Archive. While existing research largely focuses on English-language forums dominated by North American Incels (and to a lesser extent, European), our work addresses this gap by creating a multilingual and multimodal corpus from the Italian and English Incel forums. The Incelosphere, comprising forums, blogs, and websites, serves as a cross-cultural case study of male supremacist communities. Our contribution lies in offering an original cross-cultural perspective on incels and discussing challenges in constructing a multimodal and multilingual corpus, which preserves the linear and conversational structure of the forum. 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 using a novel topic modeling technique based on Transformer architectures. Results reveal differences in discussion topics and in the targets of hate between Anglophone and Italian communities. The Anglophone incel community displays frequent marks of anti-Semitic and racist discourses, associating Incel identity with perceived social issues among non-white users. Conversely, the Italian forum exhibits less emphasis on such trends, with stereotypes and discrimination focusing on regional distinctions (so called *antimeridionalism*) rather than immigration. This disparity represents a point of divergence between the two communities and may offer valuable insights for future analyses aimed at deepening the cultural context of the user base and their radical expressions.

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

https://doi.org/10.1515/9783111434018-012

Keywords: Computer Mediated Communication, cross-cultural analysis, Incels, online extremism, topic modeling, multimodal corpora, comparable corpora

1 Introduction

In this chapter we describe the steps that led to the creation and exploratory analysis of *IDA – Incel Data Archive*, a multimodal comparable corpus of forum-based interactions of the incel community in English and Italian. For the composition of the comparable corpus, data were collected from two of the most populated incel fora in both languages: *incels.is* and *Il Forum Dei Brutti* (*Forum of the ugly people*, from now on abbreviated in FDB). As we will detail below, our contribution is twofold. Primarily, we aim to contribute to a deeper understanding of Incel communities from a cross-cultural perspective. Secondly, from a methodological standpoint, our objective is to present a framework for the construction of a corpus designed to study forum-based communities, drawing on insights from the field of Computer-Mediated Communication, Social Sciences and Corpus Linguistics. We argue that this interdisciplinary approach holds particular value within the domain of Digital Linguistics, as digital environments constitute a multifaceted nexus of technological affordances and communicative practices. Furthermore, for researchers engaged in the analysis of online discourse, it is crucial to situate linguistic performances within the broader socio-political context.

After spreading within Reddit, Incel communities gradually aggregated outside mainstream social networks, creating an independent ecosystem of forum-based communities. Recently, several studies (Gillett and Suzor 2022; Trujillo and Cresci 2022) supported the hypothesis that moderation and quarantine practices adopted by mainstream social media, such as Facebook and Instagram, 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 digitally mediated communication that exhibited similar ideological content from various perspectives, framing the phenomenon of *toxic technoculture* (Massanari 2017) in different languages and contexts. Furthermore, even though the discourse of the Incelosphere is characterised by hate speech primarily targeting women (Heritage and Koller 2020; Sugiura 2021; O'Malley et al. 2022), we argue that a corpus consisting of data from the Incelosphere is not only of interest to those studying misogynistic discourse. Indeed, it can contribute to the research community engaged in the study of digitally mediated communication more broadly by providing real examples of spontaneous conversations for the study of asynchronous interactions between users in fora strictly connected to instances of white supremacy and to new form of populist far-right (Nagle 2017; Mamié et al. 2021).

Beyond these wider aspects, this article specifically aims to respond to recent calls for the need to contextualise violent online behaviour also in non-English speaking communities (Dwyer 2017; Schoenebeck et al. 2023). The research questions that prompted the development of the IDA corpus can thus be summarised as follows:

1. Is it possible to compare the discourses of Anglophone and non-Anglophone incel communities, and from what perspective?
2. How do such communities articulate their beliefs and aggregation purposes?

Indubitably, contextual understanding of violence, coupled with a multifaceted approach that includes political, social and interactional dimensions, can assist in developing effective forms of resistance and counteraction, highlighting the role of background culture in determining how individuals may disseminate online violence and articulate their extremist narratives. Moreover, it is crucial to investigate extremist phenomena from a perspective that is both geographically and linguistically situated. As recently stressed by Vessey (2024), "the meanings that can emerge from a multilingual perspective are generally non-obvious, and perhaps as a result they tend to be overlooked" (ibid.: 7). Therefore, this work provides a resource that allows for a more nuanced and complex understanding of the spread of such ideologies and their manifestations within different cultural domains.

Finally, it is worth noting that the discursive boundaries of online spaces are permeable, allowing content from certain web niches to infiltrate more controlled and secure spaces, such as those offered by mainstream social media. The result is that instances of violence generated by subcultures on the Internet are becoming accessible to the vast majority of people who are exposed to them. Thus, digital linguistics studies can provide a useful point of view for overcoming the limitations of the media's often incomplete and inaccurate representation of online subcultures (Heritage 2023).

2 The Incelosphere so far: a transnational ecosystem

Anglophone Incel communities have been studied from a wide variety of perspectives, ranging from psychology to sociology and discourse analysis. Many of these studies were focused on Reddit groups, such as *r/ForeverAlone* and *r/Incels subred-*

dits, which are archived in datasets and can be easily used as corpora. Therefore, given the availability of resources in the English language, our current understanding of this community primarily revolves around the Anglophone context, and particularly that of the United States. In the North American context, the academic scholarship has produced a multitude of studies related to the broader concept of the Manosphere. Here, the Manosphere is defined as an umbrella-term used by scholars to delineate a vast range of realities that share the same vision of heterosexual hegemonic masculinity (Lilly 2016; Ribeiro et al 2021), of which the Incelosphera is the most extreme of the groups. To a lesser extent, other Manosphere groups have also received attention from the academic community studying online misogyny, such as PUAs (Pick Up Artists), MRAs (Men's Right Activists) and MGTOWs (Men Going Their Own Way). The scholarly attention towards the Incels arises from the urgent need to comprehend and contextualize instances of real-world violence perpetrated by members of the community, such as the 2014 Isla Vista massacre and the 2018 Toronto Van Attack. This focus reflects a significant effort to understand the social and psychological dynamics within these groups and their impacts on society.

In Sociology, studies focused on the discursive practices, rhetoric and argumentation style, symbolism, and sexual imagery of Incel communities (Massanari 2017; Wasniewska 2020; Tranchese and Sugiura 2021; Aiston 2023; Prazmo 2022), in-group and out-group identity construction (Ging 2019; Chang 2022; Thorburn et al. 2023; Scotto di Carlo 2022), 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 2022; Heritage and Koller 2020). Empirical analyses and terrorism studies have sought to trace, also through dynamic cross-platform approaches, the spreading of violent extremism in the main Anglophone Incel communities (Ribeiro et al. 2021; Baele et al. 2021, 2023; Heritage 2023), as well as their misogynistic stances (Lilly 2016; Farrell et al. 2019).

For Baele and colleagues, "incel discourse demonstrates typical markers of extremist language: an essentialist categorization of society into sharply delineated ingroups and outgroups where the latter are linguistically dehumanized, and a conspiratorial narrative presenting the ingroup as the victim of an all-powerful structure of oppression" (Baele et al. 2023: 383–384). Moreover, although incels construct clear boundaries to define ingroups and outgroups within their discourses in relation to different social actors, "the language used towards different groups is more complex than 'in-group evaluation is good; out-group evaluation is bad'. Incels do often represent some out-group members in positive ways, and also do construct in-group members as undesirable" (Heritage 2023: 201).

The interest of academic research in understanding how incels represent social actors is not trivial and reflects a more general interest in the discursive aspects underlying the formation of groups aggregated on the basis of extremist ideologies with transnational boundaries. Indeed, this kind of research can even reveal deep cultural continuities and fractures in how incel ideology is framed, perceived, and promoted by users. Such differences are found not only where cultural discrepancies are most evident, between the Global East and West and Global North and South, but also between countries that are considered homogeneous because they are generally associated with the industrialized West.

In terms of incel identity formation, the link between the construction of masculinity and femininity is also largely determined by the culture to which the subject belongs. This was made very clear in studies of gender and racial identity construction by Heritage (2023), who recognizes a link between the continuous and frequent intertextual references made by users and U.S. culture. Indeed, according to the author, although the Incels Wiki[1] points to the existence of several incel and related fora in France, Italy, Germany, Japan, Taiwan, Turkey, Russia, and Poland, the majority of users within the English-speaking community can be identified as coming from North America and are often associated with related white supremacist ideologies.

Looking more closely at the European landscape outside of the Anglophone context, however, we see a thriving ecosystem that is highly diversified on both a linguistic and ideological basis. Indeed, the Incels Wiki notes that many of these international communities maintain beliefs and terminologies that differ from those found within the Anglophone Incelosphere. In this regard, some studies have already begun to analyse local aspects of Incel slang and identity construction in relation to users' nationality of origin. For example, Voroshilova and Pesterev (2021) analyse the Russian community by highlighting some key differences between Russian and Western spaces, which were found to be less hostile and more welcoming to women than other English-speaking Incel spaces. The same conclusions were reached for Italian spaces, which have not been extensively studied. The few studies on Italian incels have shown that the community often revolves around pop icons like Joker (Capalbi 2021), but also figures from the Italian literary canon like poet Giacomo Leopardi. De Gasperis's (2021) study, for example, emphasizes the connection between the collective imagination of Italian literature and the processes of masculine identity definition within the most popular Italian forum. Anastasi (2022) showed that Italian and European communities in general, in contrast to Anglophone and U.S. communities, are open to welcoming individuals who identify as women. However, the latter are welcomed as actual members of the communities when they are judged to be functional in perpetuating and reinforcing sexist

[1] https://incels.wiki/w/Main_Page (last accessed 14 February 2025).

and heteronormative discourses. In Germany, users of the main forum define themselves as *absolute beginners* (AB), after a song by David Bowie. In the main German community, the acronym "AB" is used to describe people who are involuntarily single, or who were deprived of sexual activity and romantic experiences until adulthood. The main forum is also free of explicitly misogynistic content, and its netiquette advises tolerant behaviour and peaceful discussion. According to Brzuszkiewicz (2020), who analysed a Francophone incel forum, French users also exhibit the same characteristics. Finally, to the best of our knowledge, few studies have investigated visual and multimodal aspects of communication within Incels' communities, such as the creation and circulation of images, videos and Internet memes (Aulia and Rosida 2022).

Despite the apparent pluralism and inclusiveness of the European Incelosphere, in our view it would be a mistake to underestimate the potential for violence within these communities. The use of language, strategies for communicating hateful content, acceptance of essentialist instances, underlying prejudice and stereotyping that contribute to discriminatory and violent attitudes, can vary based on the cultural roots of the social actors involved. In agreement with Czerwinsky's (2023) critical review, we believe that it is necessary to investigate these differences more thoroughly. This requires the development of appropriate resources for studying country-specific, non-English-speaking groups. The construction of comparable, multilingual, and accessible language resources such as IDA is therefore a fundamental first step in this direction.

3 Online discussion fora and CMC

The Computer-Mediated Communication (CMC) approach traditionally concerns the study 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 multimodality, as well as stylistic, stylometric and pragmatic elements that could provide useful insights into the process of meaning-making at the level below the utterance. Additionally, this approach distinguishes itself from other approaches to the analysis of online discourses by considering the importance of platform-specific affordances and how they can shape interaction, an aspect we aimed to preserve in the development of IDA. Indeed, forum-mediated conversations are not simply digital versions of every-day conversations, but rather represent a distinct genre, with its conditions of production and interpretation. For example, non-synchronous digital interaction promotes the presence of complex sequential organizations, with connections to previous levels 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.

Considering these aspects, the specificity of IDA should not come as a surprise. In recent years, with the advent of numerous novel social networking platforms, such as TikTok and Twitch, there has been a notable expansion in the creation of digital language corpora, designed to examine specific discourse phenomena such as those pertaining to the anti-vaccine movements, fake news, and conspiracy theories (Miani et al. 2021). As for the corpora of computer-mediated communication that are partly or wholly composed of fora, efforts are still limited.[2] They include The Mixed Corpus: New Media in Estonian[3]; the SFNET Corpus (Tuominen et al. 2003) and the Suomi 24 Corpus[4] in Finnish; the LITIS v.1 corpus[5] in Lithuanian, the Janes corpora 1.0 (Fišer et al. 2018) in Slovenian, the CoMeRe repository[6] in French (as listed in Frey et al. 2020).

Due to the lack of systematicity in the literature devoted to fora, Holtz et al.'s (2012) study was a useful starting point for understanding both the specifics of forum-mediated interaction and the reasons why fora are particularly attractive virtual venues for extremist communities. Indeed, "in fora for radical, extremist or other ideologically sensitive communities, users will express their opinions more freely and may be less concerned about social desirability" (Holtz et al. 2012: 4). Moreover, this work has guided us in understanding why, despite the advent of technologically sophisticated mainstream social media, fora are still a useful source of data collection, particularly for building linguistic resources to study online hate speech. Among the most obvious reasons, one is the almost unlimited amount of data from spontaneous conversations, which are not subject to social constraints and/or strict moderation policies. Second, the hierarchical organization into sections that allows for simplified and rapid selection of specific sections related to content of interest. However, the collection of data from discussion fora, as well as the step to create corpora that can be considered comparable, are not without prac-

[2] For the mapping of the corpora, I used the repository offered by the CLARIN infrastructure (available at the link: https://www.clarin.eu/resource-families/cmc-corpora, last accessed 14 February 2025). Not in all cases the authors of the corpus are cited, and in some cases the construction of the corpus takes place over time and involves many researchers. For this reason, in some cases I have preferred to provide direct links to the resource rather than citing a scholarly article.

[3] https://www.cl.ut.ee/korpused/segakorpus/ (last accessed 14 February 2025).

[4] http://urn.fi/urn:nbn:fi:lb-2017021502 (last accessed 14 February 2025).

[5] http://hdl.handle.net/20.500.11821/11 (last accessed 14 February 2025).

[6] https://repository.ortolang.fr/api/content/comere/v3.3/comere.html (last accessed 14 February 2025).

tical challenges. We discuss this process in the next section, providing description of the criteria used to select the two fora, the corpus design and the data collection criteria.

4 Corpus construction

4.1 Methodology for comparability

Traditionally, the construction of bilingual and multilingual corpora can be distinguished into two sub-categories, depending on the purposes, the characteristics of the texts, and the design methods. A first type of multilingual corpora are *parallel corpora*, mainly used in translation studies, facilitating a variety of interlingual comparisons. A second type of multilingual corpora, to be distinguished from the first, are *comparable corpora*.

We can say a corpus is a comparable corpus if its components or subcorpora are collected using the same sampling frame and similar balance and representativeness (McEnery and Xiao 2007). These are corpora that are not translations of texts in one or more languages, but rather texts of genres belonging to the same domains in the same time range, in different languages. In the case of comparable corpora, their juxtaposition depends on a number of complex factors concerning the frame of reference and the sampling criteria for collecting similar texts. Despite the introduction of statistical measures to determine similarity between corpora of different languages, there is no consensus within the community as to the most accurate statistical method (Sharoff et al. 2013). Moreover, López Arroyo (2020) note that comparability should consider the nature of the purposes for which the comparable corpus is designed. Considering these, McEnery and Xiao (2007) identified three criteria that need to be met to ensure comparability between two or more corpora:

1. Same genre and domain.
2. Same period of time.
3. Same participants or community.

To these criteria, López Arroyo (2020) adds three others:

4. Same format.
5. Same style.
6. Same content or topic.

Considering the last prerequisite, a problem with large comparable corpora is that we cannot always know the content of the texts *a priori*. This problem is particularly acute when trying to collect large comparable corpora from the Web using data mining techniques. To address and overcome this challenge, and to assess the comparability of corpora of unknown composition, one possibility is to apply unsupervised clustering methods such as topic modelling. (Sharoff 2013). Thus, topic modelling was applied on both datasets to attest the effective comparability of the two fora in terms of content. This is mainly because, if for *incels.is* the ideological commitment is openly declared, starting from its name, for the *FDB* forum the assessment of the adherence to the incel ideology required some further steps. In addition to the application of topic modelling, a netnographic analysis (Kozinets 2015) was conducted on the site affordances and community practices.

Finally, it is perhaps useful to specify that the choice of FDB as a source of data for the Italian language depended mainly on the size of its user base, which at the time of collection was the largest in the Italian Incelosphere.

4.2 Methodology for the analysis of the affordances

Here, *affordances* are defined as the relationship between user and technology in terms of how this relationship is constrained by the features of a site that can influence and shape user behaviours (Evans et al. 2017). In a recent study, Diaz-Fernandez and Garcia-Mingo (2023) explored the concept of *affordances* in relation to an extensive nethnographic investigation of Forocoches, a prominent Spanish forum associated with the Manosphere. The authors posited the necessity of examining online communities as digital spaces where boundaries are delineated by specific digital practices and particular performances of digital masculinity. In this study, the same methodology is employed to examine the affordances of the incels.is forum and the Italian FDB in four dimensions of analysis:

1. Criteria for forum membership.
2. Access protocols.
3. Privacy control.
4. Normative tightness.

The first dimension concerns the criteria established by the moderators for granting access to a user wishing to join the forum, while the second concerns access protocols and describes the hierarchical approach implemented to access certain types of information. The third dimension consists of managing the users' personal information. Finally, analysing the fourth dimension, there might be a wide range of rules concerning what is considered as appropriate behaviour by the users.

These codes of conduct not only relate to permissible content to share but can also define the appropriate modality of communication. This last remark has a direct impact on the pragmatic as well as the sociolinguistic aspects of communication within the digital places analysed.

4.2.1 Structure of *incels.is*

The site *incels.is* is reported by the IncelWiki[7] to be the largest incel forum in the Anglophone Incelosphere and so far, together with redpill[8]-related subreddits, represents the preferred source of knowledge about the incel culture for scholars engaged in the study of this community (Heritage and Koller 2020; Heritage 2023; Chang 2022; Ging 2019; Bogetić et al. 2023; Sugiura 2021). Established in 2017 after the r/incels subreddit has been shut down, it now has more than 25,000 members and hundreds of external visitors per day.

In terms of content organization, each section may contain more than one thread about roughly one topic. Each thread can contain a minimum of one post, and conversations can proceed asynchronously, from top to bottom of the page, and on multiple pages (the hierarchical structure of the forum is summarised in Figure 1).

Threads can also be resumed several years later. This aspect can pose a challenge for forum-mediated communication analysis, as the dates of creation of a thread may also be far remote in time from the individual posts within it. Finally, within a post it is possible to invoke another member through the tagging functionality offered by the *@Nickname* and to quote texts from other posts, thus triggering a mechanism of nesting between content.

At the time of writing, the rules of *incels.is* include: an account creation system that allows relative anonymity through the use of nicknames. Only members with formally registered accounts can access private sections and participate in conversations. Over time, access to the Anglophone forum has been restricted to: men aged 18 and over who *want a romantic relationship but are unable to have one*. Women and queer individuals are explicitly excluded.

[7] https://incels.wiki/w/Incels.is (last accessed 14 February 2025).

[8] The Red Pill is defined as an ideology that opposes feminism. Adherents of this ideology idealise a past in which masculinity was expressed through physical strength, economic and sexual power, and wish for a return to such dynamics (Heritage 2023).

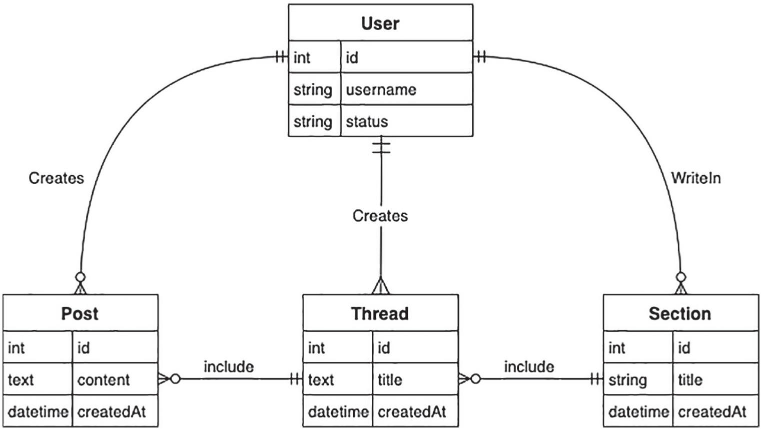


Figure 1: Fora structure diagram and associated users' metadata.

The second dimension of analysis concerns the access protocols. In fact, although most sections are public in the *incels.is* forum, others are likely to be visible and accessible only to registered users and/or users with a high hierarchical status. In addition, newcomers are not allowed to send private messages or vote in polls. These privileges are acquired over time by improving ranking and status. Status (symbolised by colourful stars) can increase on the basis of activities and longevity of the user.

On the third dimension, privacy control, incels.is invites users to take responsibility for their own privacy, for example by not publishing photos and personal information, and by connecting through proxies and VPNs. However, potentially just Admins and moderators can have access to private information about the users.

Finally, analysing the fourth dimension, there are a wide range of rules concerning what is considered appropriate behaviour by users within the sections, such as the prohibition of sharing child pornography or violent and gore-oriented content. Moreover, spamming and advertising, the expression of so-called *bluepilled* positions, sharing content that in any way supports or represents the LGBTQA+ community, posting images or content that show explicit abuse of animals, to name but a few, is prohibited. It should also be noted that it is not possible to post personal images for the purpose of receiving an aesthetic evaluation from other users, a practice which is instead the dominant theme of the community.

4.2.2 Structure of the *Forum dei Brutti*

The *Forum dei Brutti* (a possible translation in English would be Forum of the ugly people, and henceforth FDB) is the largest forum of the Italian Incelosphere, although at the time of writing it is not the only one. An element of continuity between FDB and the *incels.is* forum is represented by the homepage's organisation into sections. These sections include a *shoutbox* where temporary messages can be posted in sequence, as well as a special section devoted only to the introductions of newcomers. A section is dedicated to topics related to the Red Pill, including personal self-disclosure and stories, advice on appearance, and aesthetic evaluations of ingroups and outgroups. Additionally, there is a private area comprising three sections: one for mutual psychological support, one for discussing conspiracy theories, and one for exchanging pornographic material. Finally, a section entitled Off Topic is provided, in which users may discuss television series and other topics not related to the Red Pill ideology. This section generally hosts topics about national and foreign politics.

Considering the first dimension, formally, membership in the Italian community is open to all, regardless of gender or sexual orientation. The criteria for FDB membership are less rigid than that of *incels.is*, and more subject to case-by-case evaluation.

However, in order to participate in the discussion sections of FDB, an access protocol has been developed by moderators. In the first place, to be able to participate in conversations, every newcomer is asked to submit a detailed introduction and provide a rationale for joining the group; second, female users only have to identify themselves by posting personal data (personal picture or a voice message in which specific words are clearly pronounced) in order to verify the offline identity of the person. This is done in a special private section of the forum, accessible only upon authorisation by the admins.

With regard to the third dimension, at the time of data collection, there is a scarce control over users' privacy. For example, in a forum section with no restricted access, members of the group regularly post images of themselves in order to obtain feedback and aesthetic evaluation. The practice of aesthetic evaluation is in fact very much felt within the Italian community, in contrast to incels.is, where moderation explicitly invite users to avoid posting selfies and other sensitive material. Finally, considering the fourth dimension of analysis, the code of conduct comprises the following prohibitions: blasphemy, racism, violence against women or paedophilia; the posting of explicit violent and pornographic visual content is permitted only within the private section of the forum; incitement to suicide or self-harm, the use of demoralising phrases and insulting the moderation is forbidden.

The main affordances of the two fora and the corresponding levels of analysis considered are summarised below.

Table 1: Summary of the affordances of the two fora.

	incels.is	***FDB***
Membership criteria	Only male heterosexual users	Formally open to all gender and sexual orientation
Access protocols	Subscription and status acquisition	Mandatory introduction Off-line identity verification (only for female users)
Privacy control	Privacy control towards in-groups and out-groups Responsibility of the users	No clear rules Privacy control towards out-groups
Normative tightness	Code of conduct	Code of conduct

4.3 Collection and design rationale

For the dataset collection, we considered that both fora 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. Worth noting, we took in consideration only the public sections of the fora. For the Italian forum, we collected data from: 1. the presentation section; 2. the section dedicated to discussion about the incel condition; 3. two sections dedicated to aesthetic evaluations, 4. a section called off topic, where users can discuss topics that are not related to the Incel ideology. For the Anglophone forum, we collected data from 1. Inceldom discussions, 2. gaming, entertaining and lifestyle; 3. two sections related to off topic discussions, such as politics, philosophy and religion. We also chose to collect images, video and other multimodal material embedded in the posts, users' avatars and emoticons. The metadata collected were related to users' *nicknames*, *posting time*, *title*, *permalink*, *date* and *id* for threads and *speaker*, *content*, *permalink*, *date*, *id*, *thread id*, *title*, *image urls* and *reply to* for posts. Additionally, we decided to deal with *quotes* and *replies-to* because part of the future goal of this project is to analyze the flow of the asynchronous conversations between users. Thus, we opted to automatically tag these elements in post-processing.

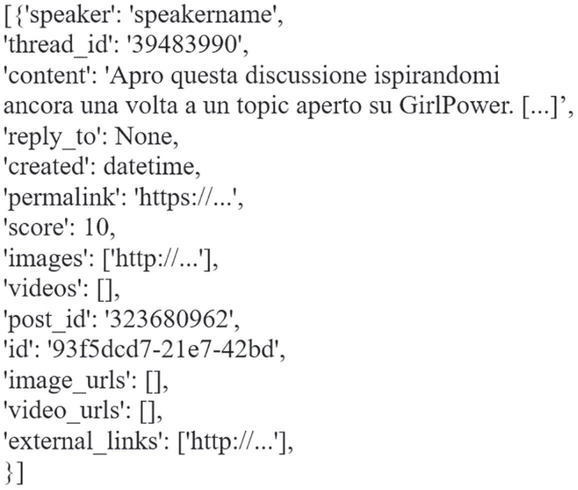
To collect the data, we implemented multiple crawlers, one for each forum, in order to systematically download threads and posts from the very beginning of the platform until March 2023, when we ended the process.

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 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.

With this procedure, the corpus is organized to capture the hierarchical structure of the fora of sections, threads and posts as well as the conversational flow of the threads and posts of referring, quoting and replying to other users. The entire content of each threads is organized as in a list of python dictionaries that associate each key with a corresponding value, as is exemplified below. For each dictionary, keys are *speaker* of each post in the thread, the *thread id* and the *post id*, *data and time* of publication, the *text*, *number of like* of the post, *images*, *videos* and *external links* (where present).

Visually, the data are organized as follow:



```
[{'speaker': 'speakername',
'thread_id': '39483990',
'content': 'Apro questa discussione ispirandomi
ancora una volta a un topic aperto su GirlPower. [...]’,
'reply_to': None,
'created': datetime,
'permalink': 'https://...',
'score': 10,
'images': ['http://...'],
'videos': [],
'post_id': '323680962',
'id': '93f5dcd7-21e7-42bd',
'image_urls': [],
'video_urls': [],
'external_links': ['http://...'],
}]
```

This organisation of data in the key-value structure allows easy retrieval of the information contained within each thread.

The crawlers for data extraction are implemented with Scrapy[9], a Python framework for extracting data from websites. To navigate the for a 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 specific to every website and forum, which makes the development of such crawlers a complex and time-consuming endeavour.

Table 2: Composition of the Anglophone and Italian datasets.

	incels.is	***FDB***
Forum sections	3	5
Number of threads	369,174	35,624
Number of posts per threads	7,359,727	740,278
Average post per thread	20	21
Average post lengths (in chars)	161.45	281.90
Images (unique)	425,259	20,183
Time span	November 2017 – March 2023	April 2009 – March 2023
Number of users (at the time of collection)	12,584	7,010

As can be observed from Table 2, despite the Italian dataset is older, it is markedly smaller in size than the Anglophone one, comprising approximately 10% of the total.

[9] https://www.sbert.net/docs/pretrained_models.html (last accessed 14 February 2025).

5 Methodology for the analysis of textual content

5.1 Transformer-based topic modeling

As Hotlz et al. (2012) points out, Internet fora are often analysed using qualitative methods. Common qualitative methods include content analysis, conversation and discourse analysis, and thematic analysis. However, having to compare two large datasets, as mentioned in section 4.1, the exploration of the content relied on topic modelling and then supporting our interpretation by reading the concordance lines (Baker and Egbert 2016).

Regarding the topic modeling, given the disproportion of the two datasets, we randomly sampled 10% of the threads from the incels.is forum (36,917), balancing the smallest FDB (35,624). Although criticised as being non-scientific method (Brookes and McEnery 2019; McEnery and Brezina 2022), in Social Sciences and Digital Humanities, topic modeling is still a widely used technique for exploring large unlabelled corpora (Jaworska and Nanda 2021). One of the main reasons for this criticism lie in the subjectivity and low intersubjective verifiability in selecting the topics to be studied. In fact, it is usually the analyst who decides how to interpret the list of words associated by the algorithm at the output stage, discarding topics that she considers wrong.

To address these criticisms, in our analysis, we replaced the LDA approach (Blei et al. 2003), based on bag-of-words representations, with a new approach based on a transformer architecture (Vaswani 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 meaningful word lists and minimized output manipulation, such as the ex-post removal of topics considered uninformative. To perform the transformer-based Topic Modeling, we used the Sentence Transformers model all-mpnet-base-v2[10] (Reimers and Gureyych 2019) to compute vector representations of the threads. Topic modeling allowed us to obtain some preliminary insights on the topic trends, both synchronically and diachronically (see Figure 2 and 3).

[10] List of pretrained sentence transformers models: https://www.sbert.net/docs/pretrainedmodels.html (last accessed 14 February 2025).

While the static visualisation of topics allowed us to obtain a general overview of the types of discourses addressed by users in the two fora, as pointed out by Jaworska and Nanda (2021), the visual representation of topics over time is an effective method to clarify the narrative dynamics or changes of topics over time. This approach has made it possible to demonstrate several trends indicative of evolving practices.

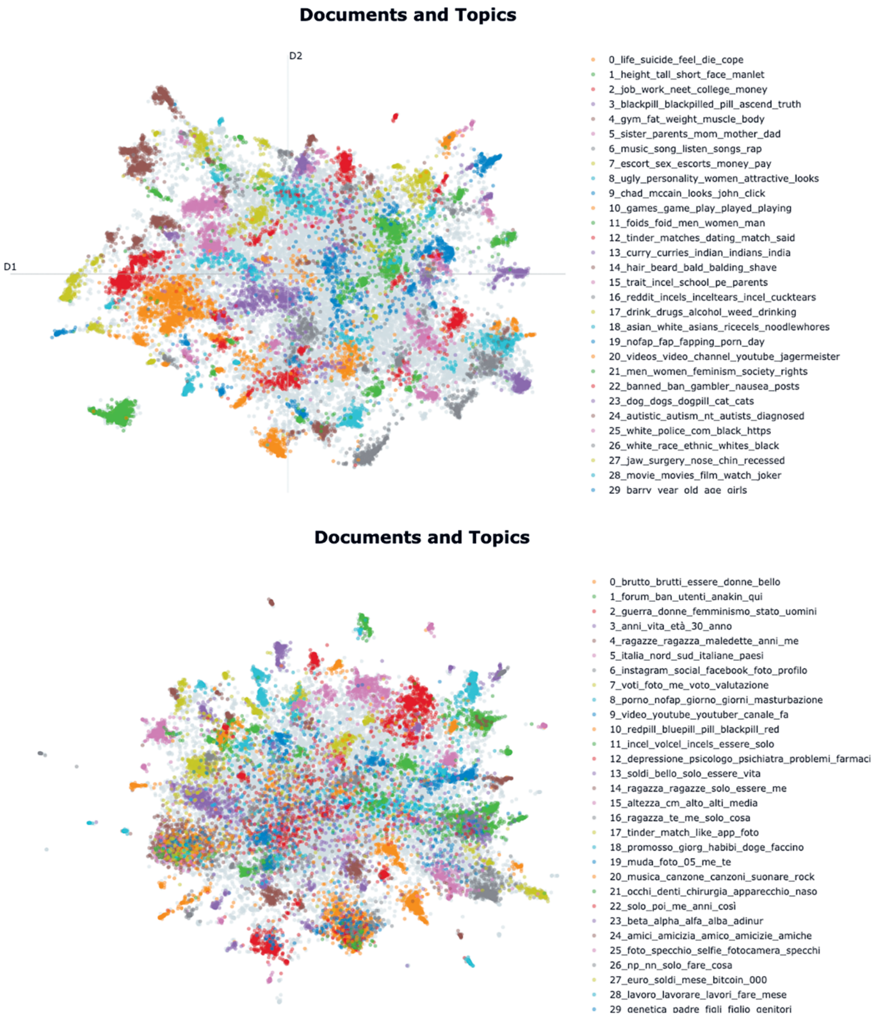


Figure 2: Static topic modelling comparison of the Anglophone (Top) and Italian (Bottom) fora. Best viewed with colour and zoom.

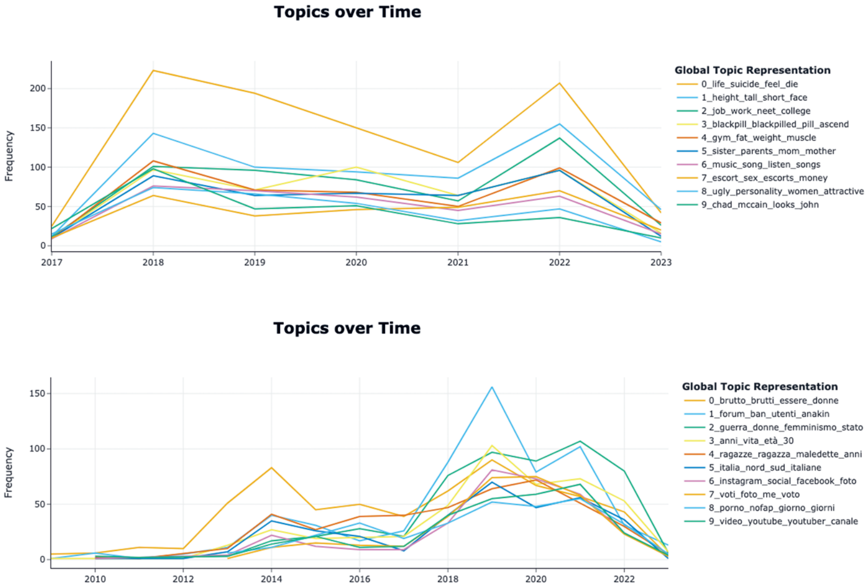


Figure 3: Dynamic topic modelling comparison of the Anglophone (Top) and Italian (Bottom) fora. Best viewed with colour and zoom.

For the purposes of this chapter, we have chosen to report only the first 30 topics generated by the static visualisation for both datasets.

5.2 Concordances

Concordance analysis consists of the in-depth study of words within a corpus, examining the immediate context in which they appear. Concordances can be organised alphabetically or according to grammatical categories, such as nouns or verbs that precede or follow a specific word, to highlight recurring patterns (Tranchese 2023). This method is useful for identifying language patterns that would otherwise be difficult to detect.

Since it is not possible to report all derived concordance lines for a single term in a large corpus, Sinclair (1999) suggests selecting random samples of concordances to identify and confirm patterns. In our case, SketchEngine's *random sample* function was used to select the concordances lines. Concordance analysis is crucial because it integrates quantitative analysis with a qualitative perspective, contextualising the results. This is one of the classical approaches in corpus linguistics, as it allows single terms to be studied in their contexts.

5.3 Overall interpretation and content comparison

So far, we have attested the comparability of the two fora from a qualitative point of view by comparing the main functionalities of the two platforms, their internal organisation, user management and rules. The results of the dynamic topic modelling (see Figure 3) offer preliminary insight into the narrative shift in both fora over the years. In the incels.is community, the top 10 topics remain stable. In particular, Topic 0 related to mental health (*life_suicide_feel_die*) seems to be significantly more discussed in the community compared to others. Furthermore, in the *incels.is* community, the most frequent topics over time are those related to aesthetics (Topic 1, Topic 4 and Topic 8). The Italian forum, on the other hand, shows a noticeable shift in user interest, with 2017 marking a crucial point in this trend. Before 2017, the dominant topic of discussion was the identity traits that characterised the user base and gave the group its name (being ugly, Topic 0). After 2017 (notably, corresponding to the opening of the incels.is forum), the focus of the community seems to have shifted to discussions about maintaining community boundaries and the internal life of the forum, as evidenced by the presence of keywords such as *ban* and *users* (Topic 1). The latter concern was assessed in the light of the development of internal forum netiquette, which saw an increase in the registration of new users as a result of the growing popularity of the Italian forum in the media.

The static clustering of the two datasets shows the top 30 general topics. These are summarised in Table 3 and each topic is associated to a semantic category.

As it can be observed from the summary Table 3, all the semantic categories identified have a counterpart in both languages, with the exception of the category *Animals.* In some cases, such as for the categories Off-line entertainment, Women and Men, and Ethnicity and Nationality, the concentration of topics is much higher in one of the two fora. This might indicate either the centrality and importance of the topic within the fora, or the interconnectedness of the topic with other discussion topics.

Table 3: The most frequent topics in the IDA corpus (summarisation of the static topic modeling).

Semantic category	incels.is topics	FDB topics
Mental health	**Topic 0**: life_suicide_feel_die_cope	**Topic 12**: depressione_psicologo_ psichiatra_problemi_farmaci
Mental health	**Topic 24**: autistic_autism_nt_autists_ diagnosed.	
Aesthetics	**Topic 1**: height, tall, short, face, manlet.	**Topic 0**: brutto, brutti, essere, donne, bello.
Aesthetics	**Topic 4**: gym, fat, weight, muscle, body.	**Topic 7**: voti, foto, me, voto, valutazione.
Aesthetics	**Topic 8**: ugly, personality, women, attractive, looks.	**Topic 15**: altezza, cm, alto, alti, media.
Aesthetics	**Topic 14**: hair, beard, bald, balding, shave.	**Topic 21**: occhi, denti, chirurgia, apparecchio, naso.
Aesthetics	**Topic 27**: jaw, surgery, nose, chin, recessed.	**Topic 25**: foto, specchio, selfie, fotocamera, specchi.
Economic status and Employment	**Topic 2**: job, work, neet, college, money.	**Topic 13**: soldi, bello, solo, essere, vita.
Economic status and Employment		**Topic 27**: euro, soldi, mese, bitcoin, 000.
Economic status and Employment		**Topic 28**: lavoro, lavorare, lavori, fare, mese.
Affective sphere	**Topic 5**: sister, parents, mom, mother, dad.	**Topic 29**: genetica, padre, figli, figlio, genitori.
Affective sphere		**Topic 24**: amici, amicizia, amico, amicizie, amiche.
Sexuality	**Topic 7**: escort, sex, escorts, money, pay.	**Topic 8**: porno, nofap, giorno, giorni, masturbazione.
Sexuality	**Topic 19**: nofap, fap, fapping, porn, day.	
Off-line entertainment	**Topic 6**: music, song, listen, songs, rap.	**Topic 20**: musica, canzone, canzoni, suonare, rock.
Off-line entertainment	**Topic 10**: games, game, play, played, playing.	
	Topic 28: movie, movies, film, watch, joker.	
Off-line entertainment	**Topic 17**: drink, drugs, alcohol, weed, drinking.	

Semantic category	incels.is topics	FDB topics
Social media and web application	**Topic 20**: videos, video, channel, youtube, jegermeister.	**Topic 6**: Instagram, social, facebook, foto, profile.
Social media and web application	**Topic 12**: tinder, matches, dating, match, said.	**Topic 9**: video, youtube, youtuber, canale, fa.
Social media and web application	**Topic 16**: reddit, incels, inceltears, incel, cucktears.	**Topic 17**: tinder, match, like, app, foto.
Forum life	**Topic 22**: banned, ban, gambler, nausea, posts.	**Topic 1**: forum, ban, utenti, Anakin, qui. **Topic 23**: beta, alpha, alfa, alba adinur.
Women and Men	**Topic 21**: men, women, feminism, society, rights.	**Topic 2**: Guerra, donne, femminismo, stato, uomini.
Women and Men	**Topic 11**: foids, foid, men, women, man.	**Topic 4**: ragazze, ragazza, maledette, anni, me.
Women and Men		**Topic 14**: ragazza, ragazze, solo, essere, me.
Women and Men		**Topic 16**: ragazza, te, me, solo, cosa.
Women and Men		**Topic 26**: np, nn, solo, fare, cosa.
Ethnicity and Nationalities	**Topic 13**: curry, curries, Indian, Indians, india.	**Topic 5:** Italia, nord, sud, italiane, paesi.
Ethnicity and Nationalities	**Topic 18**: Asian, white, Asians, ricecels, noodlewhores.	
Ethnicity and Nationalities	**Topic 26**: white, race, ethnic, whites, black.	
Incel ideology	**Topic 3**: blackpill, blackpilled, pill, ascend, truth.	**Topic 10**: redpill, bluepill, pill, blackpill, red. **Topic 11**: incel, volce, incels, essere, solo.
Age	**Topic 29**: barry, year, old, age, girls.	**Topic 3**: anni, vita, età, 30, anno.
Animals	**Topic 23**: dog, dogs, dogpill, cat, cats.	–
Unclear	**Topic 9**: chad, mccain, looks, john, click.	**Topic 18**: promosso, giorg, habibi, doge, faccino.
Unclear	**Topic 15**: trait, incel, school, pe, parents.	**Topic 19**: muda, foto, 05, me, te.
Unclear	**Topic 25**: white, police, com, black, https.	**Topic 22**: solo, poi, me, anni, così.

The Aesthetics category was found to contain the highest number of topics in both languages. References to facial features (*occhi, denti, naso* in Italian, *nose, jaw* and *chin* in English) and height (*altezza, cm, alto* in Italian e *tall, short* in English) are prominent. Below are some examples of in-context usage for terms related to aesthetics in both languages:

Concordances related to Aesthetic category in Italian:[11]

1. Cmq 176 al nord mica è **alto** siamo seri, solo al sud sei normale intorno a 175. *[176 in the north is not tall let's be serious, only in the south are you normal around 175.]*
2. Il suo difetto principale è il **naso**, e direi anche gli **occhi** ma questo è per via degli occhiali. *[His main flaw is his nose, and I would say also his eyes, but that's because of the glasses.]*
3. Secondo me con un taglio di capelli diverso e una piccola aggiustatina ai **denti** guadagna qualche punto. *[In my opinion with a different hairstyle and a little adjustment of the teeth he gains a few points.]*

Concordances related to Aesthetic category in English:

4. Just kill everyone **taller** than you.
5. I had an extremely recessed **jawline** and no **chin**.
6. In addition, curries have longer, thinner **noses** which makes their appearance sharper.

This confirms the fundamental tenets of the so-called *LMS* theory, an acronym for Look, Money, and Status. This theory posits that both men and women are regarded as, and perceive themselves to be, "sexual objects to be evaluated and placed in a hierarchical order characterised mainly by aesthetics and economic status" (Dordoni and Magaraggia 2021: 46, *our translation*). For incels, being aesthetically attractive is a value that is attributed to the subject by genetic factors and is judged in an objective way. As such, it can be measured within a numerical range from 1 to 10 (the so-called decile scale),[12] based on certain physical characteristics (i.e. bone structure, height, jaw). In these fora users often lament their physical appearance and being rejected because of it, blaming women for being superficial, materialistic and *hypergamous.*[13] This is particularly true for users of the Italian community, which have made being *brutti veri* (truly uglies) their identity claim.

Other important shared topics concern users' relationship with social media platforms and web applications. In particular, references to dating apps such as Tinder are present in both languages. Recent transdisciplinary studies have highlighted the prominent role of YouTube in disseminating and reinforcing ideologies related to both, the Manosphere and heteronormative and toxic masculinity (Papadamou et al. 2021; Mamié et al. 2021; Champion and Frank 2021, Sugiura

[11] Translations from Italian to English are provided by the authors.

[12] https://incels.wiki/w/Decile (last accessed 14 February 2025).

[13] https://incels.wiki/w/Hypergamy (last accessed 14 February 2025).

2021). Indeed, Tinder appears to play a significant role in both fora in shaping users' affective and sexual imaginaries of the opposite sex.

Concordances related to social media and applications category in Italian:

1. Però da quello che ho capito è che Meetic è un sito per incontri più sobri, mentre **Tinder** è praticamente più a sfondo sessuale. *[As far as I had understood Meetic is a site more akin to serious meetups, while Tinder is basically more sexual]*
2. Piuttosto mi rattrista questa cosa perché io mi ero illuso di trovare una relazione su **Tinder** e alla fine il livello medio di ragazze che si incontrano è questo. *[I am saddened by this because I was convinced that I would find a serious relationship on Tinder, while it turns out that the average level of girls that you meet there is this.]*
3. **Tinder** è per i bellocci da almeno 7 che sanno farsi le foto. *[Tinder is for handsome guys at least a 7 who know how to take pictures.]*

Concordances related to social media and applications category in English:

4. He is the type of guy to get 100% match rate on **Tinder.**
5. I bet the day when she arrives in Europe the first thing she'll do is install **Tinder** and fuck a white Chad.
6. On **Tinder** females can choose whoever the fuck they want.

Pornography, money and female prostitution also play an important role within the category of Sexuality. By analysing the language of the (now closed) incel communities on Reddit, Tranchese and Sugiura (2021) have traced a deep connection between incel discourses and mainstream pornography. Indeed, both discourses share a highly dehumanising vocabulary and metaphorical repertoire, even to the extent of normalising violence and rape. The stereotypical and dehumanising view of women is evident in the lexical preferences of the two communities, which have developed specific slang terms to talk about women. In fact, in the category of Women and Men we find, in addition to conventional terms such as *man*, *woman*, *girl*, slang terms such as *foid* (short form of *femoid*) and its Italian counterpart *np* (*non-person*). Analysing the representation of male and female individuals, Heritage (2023) observed that the dehumanising aspect of the use of the word foid is metaphorically derived from the mixture of the terms *female* and *android*. In the Italian community, on the other hand, the dehumanising aspect comes out of metaphor: women simply do not belong to the category of people. Finally, within the categories of Women and Men, in both languages, we find references to feminism. In fact, in both communities it is not only the biological aspect linked to the feminine that is threatened, but also the possible emancipation of women through feminism, which is to blame for the economic and social crises of contemporary society.

Concordances related to the Women and Men category in Italian:

1. Le **np** guardano un brutto con disgusto rabbioso come fossi una minaccia alla loro felicità. *[The np look at an ugly with angry disgust as I were a threat to their happiness]*
2. E io sono fiero di essere misogino, ma non solo odio le **donne,** IO ODIO ANCHE CHI NON LE ODIA MA LE RISPETTA, PERCHÈ LE **NP** DI MERDA NON RISPETTANO NOI COME UMANI. *[And I am proud of being misogynous, I do not only hate women, BUT I ALSO HATE THOSE WHO DO NOT HATE THEM BUT RESPECT THEM, BECAUSE THE SHITTY NPs DO NOT RESPECT US AS HUMANS.]*
3. E piacere ad una **NP (cioè ad un essere subumano)** non è così validante, è come avere una zecca attaccata alle palle, non vedi l'ora di liberartene. *[And to be liked by an NP (meaning a subhuman) is not that validating, it is like having a tick attached to your balls, you cannot wait to have it removed.]*

Concordances related to Women and Men and applications category in English:

4. I hate **foids** and **foids** combined with **feminism** is like pure cancer.
5. The only reason I'd pay a **foid** for is to use up and cum in her holes.
6. **Foids** can't give you this, they just use it as a tool to get stuff out of you. **Beta** for his bucks, **chad** for his semen.

In addition to these themes that emphasise, from various points of view, the centrality of sexuality within the discussions in the two fora, with often openly stereotypical and denigrating meanings, we find themes that are more akin to those of support groups. Topics such as mental health, offline entertainment, work and education, and references to the family unit were also discussed. The latter theme may also provide insight into the demographic composition of both fora, which are predominantly used by young individuals (Botto and Gottzén, 2024).

It is also noteworthy that the category of Ethnicity and Nationalities deserve particular attention. In this category, we find references to ethnic and racial groups, including both commonly used terms such as *Asian*, *white*, and *Indian*, as well as specific slang terms such as *ricecels*, *curries/curries*, and *noodlewhores*.[14]

[14] The terms *ricecel* and *curry* are used derogatorily to refer to men of different ethnicities who identify as incel (involuntarily celibate). The term noodlewhore, on the other hand, refers to women of Asian descent perceived as promiscuous.

The prominence of the theme of Nationality and Ethnicity within the Anglophone community is indicated not only by the greater number of topics associated with this category, but also by the linguistic creativity observed. This aspect does not appear to be as prominent within the Italian community. In contrast, within the Italian community, the rhetoric of supremacism addresses the distinction between men and women in southern and northern Italy. This aspect represents a point of partial continuity between the two communities and may provide interesting clues for future analyses aimed at revealing the mutual influence and adaptation of Anglophone incel discourses by peripheral communities such as the Italian one. In particular, the relationship between racist instances and white supremacy is adapted and contextualised in the Italian scenario, which is characterised by a strong north-south economic divide and an extreme stereotyping of southern Italians. This phenomenon has been documented since the beginning of Italian unification and is known as *anti-meridionalism*.

6 Conclusions and future work

In this study, we detail the processes behind the development of IDA – Incel Data Archive, as well as an initial exploration of its content. With this work, we aim to contribute to the field of digital linguistics by demonstrating how an interdisciplinary methodology combining qualitative and quantitative techniques can be used to compare corpora across languages. We also show that the development of comparable corpora requires a deep and nuanced understanding of source texts and their contextual production, emphasising the central role of the researcher's experience. In light of these consideration, we believe that digital linguistics can benefit greatly from interdisciplinary synergies, particularly those studies that integrate the analysis of computer-mediated communication with social science perspectives on the role of media in shaping human interactions. To the best of our knowledge, this represents the first comparable corpus developed from the online forums of the incel community, in multiple languages. A crucial initial step towards a deeper understanding of these transnational communities, and the threat they may pose, involves data collection and the involvement of scholars specialising in digital linguistic research outside the English-speaking world. Expanding the collection to include non-English-speaking incel communities remains a key goal of this project. Furthermore, we intend to incorporate a semantic annotation phase of IDA to capture the unique discursive features of these communities.

Although unsupervised content exploration through topic modelling has only partially revealed the potential of this dataset, we believe that the results presented here already raise important questions about cross-national influences among these groups. The computational analysis of the linguistic and non-linguistic characteristics of these communities not only offers new insights into their discursive and social structure, but also helps to delineate the dynamics of radicalisation and ideological diffusion in globalised contexts. It also suggests the need for coordinated academic and policy efforts to understand and prevent extremist phenomena developing in the shadows of digital networks.

References

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

Anastasi, Selenia. 2022. 'I am not like all the other girls'. Femcel, pinkpilled and women in the Incel communities. A qualitative analysis of the Italian 'Il Forum dei Brutti'. *28th Lavender Languages and Linguistics International Conference*. Catania, Italy, 23–25 May.

Arroyo, Belén López. 2020. Can comparable corpora be compared? *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)* 39, 43–68.

Aulia, Mahirza Putra & Ida Rosida. 2022. The phenomenon of involuntary celibates (Incels) in internet meme culture: A reflection of masculine domination. *International Journal of Media and Information Literacy* 7 (1), 4–17.

Baele, Stephane, Lewys Brace & Debbie Ging. 2023. A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence* 36 (3), 1–24.

Baker, Paul & Jesse Egbert (eds.), 2016. *Triangulating methodological approaches in corpus linguistic research.* London: Routledge.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43, 209–226.

Blei, David, Andrew Ng & Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

Bogetić, Ksenija, Frazer Heritage, Veronika Koller & Mark McGlashan. 2023. Landwhales, femoids and sub-humans: Dehumanising metaphors in incel discourse. *Metaphor and the Social World* 13 (2), 178–196.

Botto, Matteo & Lucas Gottzén. 2024. Swallowing and spitting out the red pill: Young men, vulnerability, and radicalization pathways in the manosphere. *Journal of Gender Studies* 33 (5), 596–608.

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

Brzuszkiewicz, Sara. 2020. *Incel radical milieu and external locus of control* 1. International Centre for Counter-Terrorism (ICCT).

Capalbi, Antonella. 2021. Le rappresentazioni audiovisive come strumento di indagine della manosphere. Joker, supereroe per gli Incel italiani? *AG About Gender-International Journal of Gender Studies* 10 (19), 105–130.

Champion, Amanda & Richard Frank. 2021. Exploring the "radicalization pipeline" on YouTube. *Terrorism Risk Assessment Instruments: Contemporary Policy and Law Enforcement Challenges*, 152–359.

Chang, Winnie. 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.

De Gasperis, Arianna. 2021. "Giacomino uno di noi". Letteratura italiana e pratiche di maschilità nel Forum dei Brutti. *AG About Gender-International Journal of Gender Studies* 10 (19), 68–104.

Díaz-Fernández, Silvia & Elisa García-Mingo 2022. The bar of Forocoches as a masculine online place: Affordances, masculinist digital practices and trolling. *New Media & Society* 26 (9), 5336–5358.

Dordoni, Annalisa & Sveva Magaraggia. 2021. Modelli di mascolinità nei gruppi online incel e red pill: Narrazione vittimistica di s ́e, deumanizzazione e violenza contro le donne. *AG About Gender-International Journal of Gender Studies* 10 (19), 35–67.

Egger, Roman & Yu, Joanne. 2022. A topic modeling comparison between LDA, NMF, top2vec, and BERtopic to demystify twitter posts. *Frontiers in sociology* 7.

Evans, Sandra K., Katy E. Pearce, Jessica Vitak & Jeffrey W. Treem. 2017. Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer Mediated Communication* 22 (1), 35–52.

Farrell, Tracie, Miriam Fernandez, Jakub Novotny & Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, 87–96.

Fišer, Darja, Nikola Ljubešić & Tomaž Erjavec. 2018. The Janes project: language resources and tools for Slovene user generated content, *Language Resources and Evaluation* 54 (1), 1–24.

Frey, Jennifer-Carmen, Alexander König, Egon Stemle, Achille Falaise, Darja Fišer & Harald Lüngen. 2020. The FAIR Index of CMC Corpora. In Julien Longhi & Claudia Marinica (eds.), *CMC Corpora through the prism of digital humanities*, 127–144. Paris: L'Harmattan. https://cmc-corpora.org/publications/frey-et-al-2020-fair-index-cmc/ (last accessed 25 October 2024).

Gillett, Rosalie & Nicolas Suzor. 2022. Incels on reddit: A study in social norms and decentralised moderation. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2021i0.12171.

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

Grootendorst, Maarten. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. https://arxiv.org/pdf/2203.05794 (last accessed 25 October 2024).

Heritage, Frazer & Veronika Koller. 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.

Heritage, Frazer. 2023. *Incels and ideologies: Exploring how incels use language to construct gender and race*. Cham: Springer Nature.

Herring, Susan C. & Jannis Androutsopoulos. 2015. Computer-mediated discourse 2.0. In Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds.), *The handbook of discourse analysis*, 127–151. Wiley Blackwell.

Holtz, Peter, Nicole Kronberger & Wolfgang Wagner. 2012. Analyzing internet forums: A Practical Guide. *Journal of Media Psychology* 24 (2), 55–66.

Jaworska, Sylvia & Anupam Nanda. 2018. Doing well by talking good: A topic modelling-assisted discourse study of corporate social responsibility. *Applied Linguistics* 39 (3), 373–399.

Kozinets, Robert. 2015. *Netnography: Redefined*. Los Angeles: Sage.

Lilly, Mary. 2016. *"The world is not a safe place for men": The representational politics of the Manosphere*. Doctoral dissertation, University of Ottawa.

Mamié, Robin, Manoel Horta Ribeiro & Robert West. 2021. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In *Proceedings of the 13th ACM Web Science Conference*, 139–147.

Massanari, Adrienne. 2017. #gamergate and the fappening: How reddit as algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19 (3), 329–346.

McEnery, Tony & Richard Xiao. 2007. Parallel and comparable corpora: What is happening. Incorporating corpora. *The Linguist and the Translator* 2 (5), 18–31.

Miani, Alessandro, Thomas Hills & Adrian Bangerter. 2021. LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 1–24.

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

O'Malley, Roberta Liggett, Karen Holt & Thomas J. Holt. 2022. An exploration of the involuntary celibate (incel) subculture online. *Journal of Interpersonal Violence* 37, 7–8, 4981–5008.

Papadamou, Kostantinos, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini & Michael Sirivianos. 2021. "How over is it?" Understanding the Incel Community on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 1–25.

Pelzer, Björn, Lisa Kaati, Katie Cohen & Johan Fernquist. 2021. Toxic language in online incel communities. *SN Social Sciences* 1, 1–22. https://doi.org/10.1007/s43545-021-00220-8.

Prazmo, Ewelina. 2022. In dialogue with non-humans or how women are silenced in incels' discourse. *Language and Dialogue* 12 (3), 383–406.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. https://arxiv.org/pdf/1908.10084 (last accessed 25 October 2024).

Ribeiro, Manoel Horta, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg & Savvas Zannettou. 2021. The evolution of the manosphere across the web. *In Proceedings of the International AAAI Conference on Web and Social Media* 15, 196–207.

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

Sharoff, Serge, Reinhard Rapp, Pierre Zweigenbaum & Pascale Fung (eds.), 2013. *Building and using comparable corpora*. Berlin & Heidelberg: Springer.

Sharoff, S. 2013. Measuring the distance between comparable corpora between languages. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum & Pascale Fung (eds.), *Building and using comparable corpora*, 113–130. Berlin & Heidelberg: Springer.

Sinclair, John. 1999. A way with common words. In Hilde Hasselgard & Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 157–179. Amsterdam: Rodopi.

Sugiura, Luisa. 2021. *The incel rebellion: The rise of the manosphere and the virtual war against women*. Bingley, UK: Emerald Group Publishing.

Thorburn, Joshua, Anastasia Powell & Peter Chambers. 2023. A world alone: Masculinities, humiliation and aggrieved entitlement on an incel forum. *The British Journal of Criminology* 63, 1, 238–254.

Tognini-Bonelli, Elena. 2007. The corpus-driven approach. In Wolfgang Teubert & Ramesh Krishnamurthy (eds.), *Corpus linguistics: Critical concepts in linguistics*, 74–92. London: Routledge.

Tranchese, Alessia & Luisa Sugiura. 2021. "I don't hate all women, just those stuck-up bitches": How incels and mainstream pornography speak the same extreme language of misogyny. *Violence Against Women* 27, 14, 2709–2734.

Tranchese, Alessia. 2023. *From Fritzl to #metoo*. Cham: Springer International Publishing.

Trujillo, Amaury & Stefano Cresci. S. 2022. Make reddit great again: Assessing community effects of moderation interventions on r/the_donald. In *Proceedings of the ACM on Human Computer Interaction* 6 (CSCW2), 1–28.

Tuominen, Tuuli, Panu Kalliokoski & Antti Arppe. 2003. SFNET Corpus [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-20150126 (last accessed 25 October 2024).

Vessey, Rachelle. 2024. 'From cross-linguistic to intersectional corpus-assisted discourse studies', *Journal of Corpora and Discourse Studies* 7 (1), 7–21.

Voroshilova, Anzhelika I. & Dmitriy O. Pesterev. 2021. Russian incels web community: Thematic and semantic analysis. In *2021 Communication Strategies in Digital Society Seminar* (ComSDS), 185–190.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, USA.

Waśniewska, Małgorzata. 2020. The red pill, unicorns and white knights: Cultural symbolism and conceptual metaphor in the slang of online incel communities. In Barbara Lewandowska-Tomaszczyk (ed.), *Cultural conceptualizations in language and communication*, 65–82. Cham: Springer International.