Blue-haired, misandriche, rabiata: Tracing the Connotation of 'Feminist(s)' Across Time, Languages and Domains

Arianna Muti[1], **Sara Gemelli**[2,3], **Emanuele Moscato**[1], **Emilie Francis**[4], **Amanda Cercas Curry**[5], **Flor Miriam Plaza-del-Arco**[6], **Debora Nozza**[1]
[1]Bocconi University, [2]University of Bergamo, [3]University of Pavia,
[4]University of Gothenburg [6] LIACS, Leiden University [5]CENTAI Institute
{arianna.muti, emanuele.moscato2, debora.nozza}@unibocconi.it,
sara.gemelli@unibg.it, amanda.cercas@centai.eu,
f.m.plaza.del.arco@liacs.leidenuniv.nl, emilie.francis@gu.se

Abstract

Understanding how words shift in meaning is crucial for analyzing societal attitudes. In this study, we investigate the contextual variations of the terms *feminist, feminists* along three axes: time, language, and domain. To this aim, we collect FEMME, a dataset comprising the occurrences of such target terms from 2014 to 2023 in English, Italian, and Swedish in two domains. For the general domain we consider Twitter and Reddit; for the hate domain we consider the Incel community. We use frame analysis, fine-tuning, and LLMs to find connotations of target terms. We find that *feminists* has a consistently more negative connotation than *feminist*. This finding indicates more hostility towards feminists as a collective, which often triggers greater societal pushback, reflecting broader patterns of group-based hostility and stigma. Across languages, we observe similar stereotypes towards feminists that include body shaming and accusations of hypocrisy and irrational behavior. Across time, we identify events that trigger a peak in terms of negative or positive connotation. As expected, the Incel spheres show predominantly negative connotations, while the general domains show mixed connotations.

***Warning**: this paper contains obfuscated examples some readers may find upsetting and offensive.*[1]

1 Introduction

While misogyny is understood as hatred or contempt towards women (Srivastava et al., 2017; Manne, 2017), anti-feminist hostility is frequently dismissed as a mere neutral political or ideological stance. The way anti-feminist rhetoric is framed can influence how the discourse around feminism evolves, ultimately shaping its connotations and affecting public opinion and social movements.

Over time, feminism has been reclaimed as a symbol of empowerment but also weaponized to discredit gender equality efforts, often shifting between praise and stigma. Analyzing these changes helps reveal resistance to feminist goals, the impact of media framing, and the persistence of misogynistic narratives. In media and popular culture, references to feminist movements and their supporters have often been framed in negative or mocking terms, reinforcing long-standing stereotypes about feminist women (North, 2009), which were assigned labels such as "femin*zis," "man-haters," and "bra-burning crazies" (Swirsky and Angelone, 2014). The persistence of such stereotypes about anyone identifying with the term may contribute to the reluctance of many women to self-identify as feminists (McCabe, 2005). On the other hand, the complexity of feminist movements, both in terms of their diverse schools of thought and their evolution over time (the so-called *waves*), has resulted in a wide and heterogeneous set of values associated with the term. This complexity is reflected in the perception of the term, leading to the different connotations it acquires in media usage and online communication. Despite the fact that fourth-wave feminists rely heavily on social media as their primary channel for communication and activism, online spaces continue to exhibit many of the stereotypes that emerged fifty years ago. These stereotypes, often renewed and adapted to the contemporary context, remain tied to the use of the term. As Ahmadi (2024) pointed out, the term *feminist* itself can be used as a pejorative epithet: to call someone a feminist can be perceived as an insult, or more generally, can carry negative connotations. From a computational perspective, Muti et al. (2024b) show that the terms *femminista, femministe* are used pejoratively, as a slur, in Italian tweets.

However, to the best of our knowledge, no prior work in NLP has explored the extent of this phe-

[1]Examples have been obfuscated with a Python package for obfuscating profanities: PrOf by Nozza and Hovy (2023).

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 299–311

nomenon. **This study is the first to systematically investigate how frequently the terms *feminist(s)* are used with negative connotations across different languages, time periods, and domains**. Specifically, we focus on the period from 2014 to 2023, considering three languages: Italian, English and Swedish. For the general domain, we consider Twitter for the Italian and Reddit for the English and Swedish languages, a choice based on platform usage and available data. For the hate domain, we consider Incel forums. Incels, short for *involuntary celibates*, pertain to the so-called *manosphere* (Nagle, 2017) and mainly comprise men who struggle to find a sexual partner or significant other, and blame this on women and feminists. Some members of this community tend to engage in the spread of various forms of hate speech, in particular misogyny.

We address two research questions:

RQ1 What are the stereotypes associated with *feminist(s)* across time, languages and domains?

RQ2 Are there events that trigger a shift in the connotation of the target terms *feminist, feminists* across time, for each language and domain?

2 Related Work

2.1 On Online Perception of Feminists

Several studies have focused on the ways in which feminists and feminist movements are represented and targeted in online environments. Lewis et al. (2019) examine online abuse targeting feminist women through a survey and in-depth interviews, finding clear parallels with offline gendered abuse. Dafaure (2022) analyzes the persistence of misogynistic and anti-feminist attitudes in anglophone online spaces, contextualizing them historically and showing how digital content, such as memes, YouTube videos, and social media posts, often constructs feminists as deficient in psychological, physical, or ideological terms. This aligns with the findings of Suárez Estrada et al. (2022), who examine how the affective political participation of women is monitored and disciplined in polarized online discourse surrounding feminist protests in Mexico. Their study reveals that feminist protesters were subjected to hate speech and toxicity, and that their affective agency was often silenced and perceived as inappropriate relative to socially sanctioned gendered norms, ultimately reinforcing the very stereotypes they seek to challenge. Similarly, Dickel and Evolvi (2023) investigate discussions of the #MeToo movement within two misogynistic manosphere groups, identifying several recurring themes. Among these is the portrayal of #MeToo as 'feminist propaganda', which reinforces anti-feminist discourse centered on reclaiming power through the reassertion of patriarchal norms. Focusing specifically on the manosphere, Aiston (2024) conducts a qualitative analysis of an anti-feminist Reddit community, showing that feminists are consistently depicted as a unified, misandric group seeking dominance over men.

2.2 Misogyny in NLP

Misogynistic discourse varies across online communities, often adopting specific vocabulary and norms. The incel (involuntary celibate) subculture exemplifies this, using unique jargon to express extreme anti-women views. Research shows incel forums generate significant gender-based hate, much of it in coded, community-specific language (Yoder et al., 2023). By applying large-scale frame semantic analysis, Gemelli and Minnema (2024) explore how the users of a popular Italian incel forum conceptualize the world and their experiences, and especially the way they write about gender issues, men and women. Guest et al. (2021) include a variety of manosphere-related subreddits in their dataset. The EXIST 2021 challenge (sEXism Identification in Social Networks) (Rodríguez-Sánchez et al., 2021) and the EDOS 2023 (Explainable Detection of Online Sexism) task at SemEval-2023 (Kirk et al., 2023) include anti-feminist posts in their dataset. Muti et al. (2024a) include anti-feminist data in their ImplicIT-Mis dataset containing implicit misogynistic Facebook comments in Italian. To the best of our knowledge, no multilingual dataset has been created for anti-feminist language. To fill this gap, we introduce FEMME, the first resource of its kind that includes general and hate domain anti-feminist discourse.

3 The FEMME Dataset

We collect FEMME, FEminist across Multilingual and Multidomain Eras, a multilingual dataset comprising occurrences of the terms *feminist* and *feminists* extracted from different online spaces: the hate domain includes posts from Incel communities, while the general domain comprises data from Twitter and Reddit. Such dataset ensure temporal coverage, allowing us to track the evolution of the terms across different periods, from 2014 to

Category	Description
Insult	Feminist(s) refers to a person and is characterized by adjectives or expressions with negative connotations, or it occurs with slurs. Subframes: **intellect**, **physical aspect**, or **sexuality**.
Inanimate	Feminist(s) refers to an inanimate concept and is used as an adjective to modify nouns or concepts with negative connotations.
Distance	«I am not a feminist, but...»: the users distance themselves from the movement or its values and ideas.
List	Feminist(s) is part of a list of elements perceived as negative.
Double Standard	Feminist(s) is associated with the concept of hypocrisy, often related to their behavior with men.
Stereotype	Feminist(s) is associated with stereotypical features, without direct insult. Subframes: **intellect**, **physical aspect**, or **sexuality**.
Sarcasm	The sentence conveys negativity through irony or sarcastic phrasing, often mocking feminism.
Misandry	Feminist(s) are characterized as hating men.
DARVO	"Deny, Attack, Reverse Victim and Offender". Feminists are portrayed as evil, violent, power-hungry, and destroying society. The user is framing feminists as real oppressors, flipping the narratives.
Men	The negative connotation is directed at men being feminists.
Attitude	The authors express hate or violence against feminists without expressing a connotation of them.
Intersect	Feminism is associated with other hate speech topics like racism, religion, homophobia, or right-wing ideology.
Dismissal of Feminism	Feminist causes or women's rights are dismissed as invalid, exaggerated, or nonsense.

Table 1: Frame Annotation Categories

2023. For each language, we employ the following keywords: *femminista, femministe* for Italian (IT), *feminist, feminists* for English (EN), and *feminist, feminister* for Swedish (SE).

3.1 Data Collection

For English general domain, we extract all Reddit posts[2] and comments containing the term "*feminis**" in either the body or the title for the years 2014-2023. For the hate domain, we take data from Gajo et al. (2023), a multilingual corpus for the analysis and identification of hate speech in the domain of inceldom built from incel Web forums including English.

For Swedish, we collect data from the r/Sweden forum on Reddit for the years 2014-2022 with the Pushshift API. The hate domain consists of data from Flashback retrieved with Språkbanken Text (2024) for the years 2016 to 2023. Unfortunately, the data for the Swedish hate domain is sporadic. As shown by Stenavi and Bengston (2020), the Incel community in Sweden is among the top traffic to English Incel forums. As such, there has not been a need for a Swedish language Incel forum outside of Flashback.

Domain	IT	EN	SE
General	1,050	150	1000
Hate	950	150	300
Total	2,000	300	1,300

Table 2: Statistics for annotated data.

For Italian, for the general domain we select instances from TWITA (a collection of tweets identified as being written in the Italian langauge) for the years 2014-2022 (Basile and Nissim, 2013) and Pejorativity (Muti et al., 2024b) (a corpus of misogynous tweets, containing the word *femminista/e*) for 2023. For the hate domain, we select instances from the Forum dei Brutti (FdB) corpus (Gemelli and Minnema, 2024), which includes all threads from the largest section of *Il Forum dei Brutti*, the most popular Italian incel forum, from 2010 to 2023. We only consider posts from 2014 to 2023 for consistency with other languages.

Table 4 in the Appendix shows the sources and the statistics for all data collected.

3.2 Data Annotation

By adopting a prescriptive paradigm (Rottger et al., 2022), we develop an annotation framework to capture the connotations of the terms *feminist(s)* and how they are portrayed. In line with the paradigm of Guzmán-Monteza (2023), two Italian NLP researchers who identify as feminists conducted a pilot study on 300 Italian instances, initially using two labels: negative and non-negative[3]. If the connotation is negative, the annotators specify the appropriate frames and subframes (if applicable) that best describe the type of negativity expressed (see Table 1). Starting with moderate agreement (Cohen's Kappa inter-annotator agreement of 0.7570 on the binary task and 0.52 Jaccard similarity on the multi-label), the annotators refined and updated the guidelines based on edge cases. These revised guidelines were then shared with two additional

[2]We use the Pushshift API dataset (Baumgartner et al., 2020) as a source for the posts.

[3]In this paper, non-negative connotation refers to both positive and neutral connotations.

annotators for English and one for Swedish, for which we were unable to identify a second expert native speaker. These annotators, who are native speakers of the respective languages and experts in NLP and gender-related studies, followed the same process: first annotating separately, then resolving all disagreements collaboratively. The inter-annotator agreement for the binary task in English is 0.8167 (Cohen's Kappa) and 0.26 (Jaccard similarity) in multi-label. A different number of instances were annotated for each language (see Table 2). Full guidelines and examples can be found in the Github repository.[4]

4 Frame Analysis

We use frame analysis (Entman, 1993) to identify how certain aspects in negative connotations of the terms *feminist(s)* are emphasized and to characterize evolving societal attitudes. This study has been performed on the annotated dataset. We follow a multi-step approach.

1. **Frame Identification** We first identify the semantic frames surrounding the target terms through annotation. These frames serve as an interpretive lens for understanding the social positioning of the terms feminist(s).

2. **Characterizing Words Extraction** Within each frame, we employ GPT-4o to extract representative words used to describe feminists. Characterizing words represent frames associated with feminists. The list of the characterizing words, specific to each language, is available in the Github repository.[4]

Figure 1 shows the frame frequencies by domain and language calculated over negative comments.

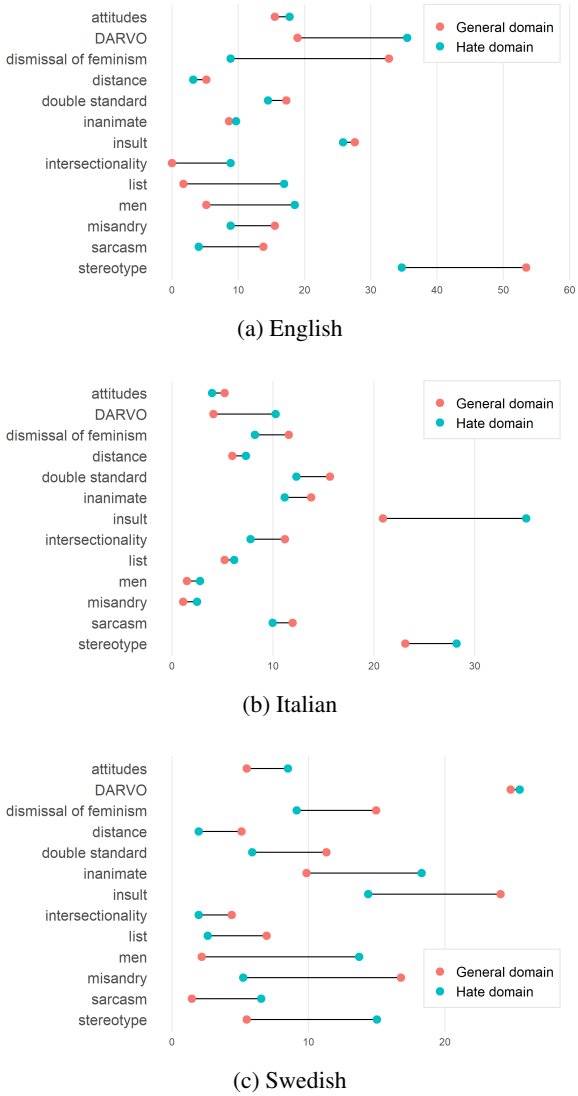
Italian Comments in which the word *femminista/e* is used with a negative connotation represent 47.4% of the data in the Italian dataset. When examining the negative comments by domain, however, we found that in the hate domain the negative comments account for 71.7%, considerably higher than the general corpus (25.5%). The most frequently used frame in the hate domain is INSULT, appearing in 35.1% of the negative comments in the dataset, associated with body shaming and politically-charged terms. This is followed by STEREOTYPE (28.2%) and DOUBLE STANDARD (12.3%). The frame INANIMATE ranks fourth (11.2%), indicating that the term *femminista/e* is often used as an adjective modifying non-human entities or abstract concepts. This is evident in the words associated with this frame, which include terms such as *ideologia* (ideology), *str*nzate* (bullshit), and *follia* (madness) (see Table Characterizing Words in the Github repository). In the general dataset, 25.5% of the comments express a negative attitude toward feminists. The three most frequent frames are the same as those in the hate domain, with STEREOTYPE being the most prevalent (23.1%), followed by INSULT (20.9%), and DOUBLE STANDARD (15.7%); notably, this last frame appears more frequently compared to the hate domain. An interestingly prominent frame in the hate domain is DARVO (Deny, Attack, Reverse Victim and Offender). In comments annotated with this frame, feminists are portrayed as individuals who position themselves as victims while actually exerting power. Alongside the DOUBLE STANDARD frame, this reinforces the stereotype of women, particularly feminists, as a homogeneous and untrustworthy group covertly intent on oppressing men. The DARVO frame appears in 4.1% of the negative comments in the general dataset, compared to 10.3% in the hate dataset. Interestingly, in the general domain the frame DISMISSAL OF FEMINISM appears more frequently (11.5%) than in the hate dataset (8.2%). This may suggest that, outside of more radicalized online spaces, anti-feminist attitudes are expressed in a less overt manner, less through direct insults and more through the discrediting of feminist struggles.

English 60% of the comments in the English dataset express negative views towards feminists. When examining the domains separately, 38.7% of the comments are negative in the general domain, compared to 82.7% in the hate domain. The general corpus shows the frame STEREOTYPE in 53.4% of negative comments, making it the most frequently annotated frame. This is significantly higher than in the hate corpus, where it appears in 34.7% of cases and ranks second after DARVO (35.5%). The third most frequent frame in the hate corpus is INSULT (25.8%). The second most frequent frame in the general corpus is DISMISSAL OF FEMINISM (32.7%), which in turn is attested only in 8.9% of the negative comments in the hate domain. This may suggest that, outside explicitly radicalized spaces, anti-feminist sentiments

[4] https://github.com/arimuti/FEMME

are more often conveyed through the delegitimization of feminist causes rather than overt hostility. Many of the typical words for this frame pertain to the semantic domain of irrationality, such as *degenerate, primitive creatures, emotionally driven, logically incapable, hysterical, inconsistent*. This indicates that the feminist causes are not only minimized, but also framed as the product of women's illogical thinking and imagination, thus dismissing their legitimacy and existence altogether. The frame DOUBLE STANDARD is also relatively frequent in the general corpus, appearing in 17.2% of negative comments (14.5% in the hate corpus) and representing feminists as *hypocrites, unreliable, brainwashed* and *misandrist* that only care about their interests, disregarding men's rights and other causes more in general. Notably, the frame MEN appears in 18.5% of negative comments in the hate dataset, versus just 5.2% in the general domain. This indicates that the negative connotations associated with the term *feminist(s)* in this community also extend to male feminists, who, although less frequently mentioned, are still consistently targeted in Incel discourse. Finally, the frame LIST appears in 16.9% of negative comments in the hate dataset, compared to just 1.7% in the Reddit corpus. Users of the Incel forum often include the term *feminist(s)* in lists alongside negatively connoted terms, thereby contributing to the construction of its pejorative meaning.

Swedish In the Swedish dataset, comments that express a negative stance towards feminists are 32.8% in total. The general dataset contains 27.4% negative comments, while in the hate domain the percentage rises to 51%. In the general domain, two frames appear most frequently within the negative comments with almost identical frequency: DARVO (24.8%) and INSULT (24.1%). These are followed by MISANDRY (16.8%) and DISMISSAL OF FEMINISM (15%) as the third and fourth most prevalent frames. The DARVO frame appears to be the most frequent one also in the hate domain, present in 25.5% of the negative comments. In the Swedish dataset, comments in the general and in the hate domain seem to express negative connotations through the same strategies in the majority of the cases. However, in the hate dataset the second, fourth, and fifth frames are respectively INANIMATE (18.3%), STEREOTYPE (15%), and INSULT (14.3%). The frequencies of these two frames are particularly interesting. While STEREOTYPE is



(a) English

(b) Italian

(c) Swedish

Figure 1: Frame frequencies by domain and language calculated over negative comments [%].

more present in the hate dataset than in the general domain (5.5%), INSULT is significantly less frequent. Finally, the frame MEN appears significantly more often in the hate domain (13.7%) than in the general domain (2.2%). This indicates that, similar to findings in the English dataset, men who identify as feminists face criticism and ostracism within the Incel community, being relegated to the out-group alongside feminists more broadly.

5 Societal Attitudes towards Feminists at Scale

In this section, we interrogate contemporary attitudes toward feminists by combining classification tools with event-driven textual analysis. We structure the experiments and the analysis around two core components: (i) connotation tracking of the

Model	IT	EN	SE
GPT-4o-mini	0.500	0.582	0.368
XLM-T	0.658	**0.920**	0.581
AI-Sweden RoBERTa	–	–	**0.700**
AlBERTo	**0.700**	–	–

Table 3: Macro F1 on binary task.

terms through binary classification (negative and non-negative) of entire posts[5] and (ii) identification of events that cause a shift in the discourse around *feminist(s)*.

5.1 Connotation prediction

Our aim is to investigate if, across years, languages, and domains, the target terms *feminist* and *feminists* undergo a connotative shift. In order to do that, we employ **encoder-based models** and **Large Language Models (LLMs)** to predict the binary label associated with the connotations, i.e. negative and non-negative. For encoder-based models, we use Twitter-XLM for English, (Barbieri et al., 2022), AI-Sweden RoBERTa[6] for Swedish, and AlBERTo for Italian (Polignano et al., 2019). All models were trained on Twitter data, except for the Swedish model, which was trained on a dataset that also includes Reddit content. For LLMs, we use GPT-4o-mini in zero-shot and few-shot settings, where one instance is reported for each category. Appendix B.2 shows the prompt.

Due to the domain sensitivity of discourse around feminism, especially in hate-prone spaces, we experiment with domain-adaptive fine-tuning. Specifically, we fine-tune models on datasets partitioned by domain. We compare this against a unified training set mixing both domains. We train all languages jointly, leveraging multilingual transfer learning to mitigate the issue of data scarcity in low-annotated languages (Röttger et al., 2022). Table 3 shows the results for each model considering the whole data, while Table 6 in Appendix D shows the results when considering domains separately. XLM-T demonstrates strong multilingual performance, particularly excelling in English (0.920, the highest). GPT-4o-mini, while competitive in English, underperforms notably in Italian and especially Swedish, suggesting potential limitations in adapting to lower-resource or less-aligned languages. Language-specific models outperform multilingual ones in their respective domains: AI-Sweden-RoBERTa leads in Swedish and AlBERTo in Italian. These results highlight the value of tailored pretraining on specific linguistic and cultural data. Results obtained by training and testing on each domain separately yielded lower average performance compared to using the combined dataset. This is likely due to the benefits of larger training data when domains are merged, which helps mitigate the limitations of the limited annotated set.

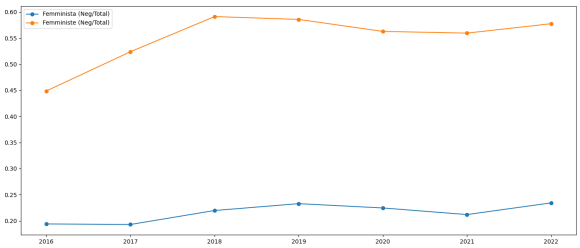
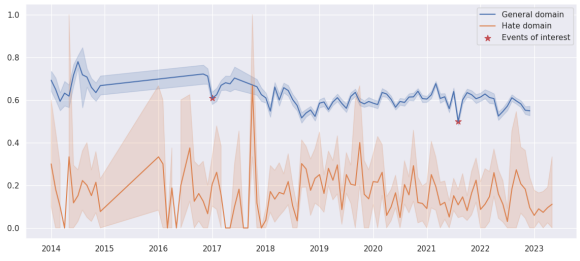


Figure 2: Comparison of the ratio of negative to total mentions for feminist and feminists in Italian.

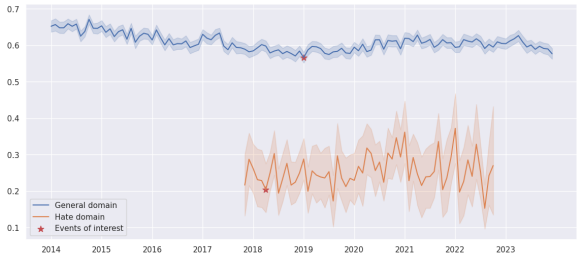
Following the results, we select the best model for each language: XLM-T for English, AlBERTo for Italian, and AI-Sweden-RoBERTa. These models are then used for prediction connotations on the whole available dataset (see Table 4). With the resulting automatically labeled large dataset, we proceed by exploring the temporal trends in how the terms feminist and feminists are perceived. Specifically, we investigate which years exhibit more negative connotations toward these terms. By aligning prediction scores with temporal metadata, we can identify periods of heightened backlash, shifts in public discourse, or events that may have influenced negative framing. This diachronic analysis allows us to situate societal attitudes within broader historical and political contexts, offering insight into how perceptions of feminists evolve over time. For Italian, Fig. 2 shows that the ratio for the plural *femministe* has a higher proportion of negative usage compared to the singular *femminista* in every observed year. While *femminista* stays in the 20% negative range, *femministe* begins near 45% negative in 2016, peaks around 59% by 2018–2019, and remains above 50% negative. This underscores a consistent and more pronounced negative framing for the plural term *femministe*. The trend is consistent in English, as can be seen in Fig. 4a in Appendix C, while it fluctuates somewhat for Swedish, as in Fig. 4b. In general, the singular

[5]We also experimented with frame prediction, but the models' performance proved inadequate.

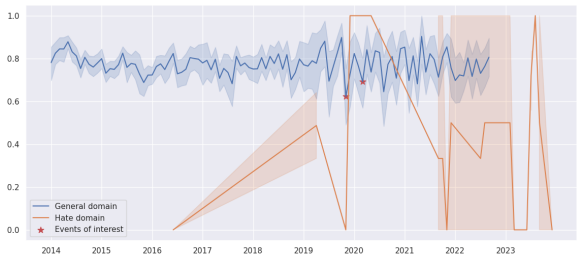
[6]https://huggingface.co/AI-Sweden-Models/roberta-large-1160k



(a) Italian



(b) English



(c) Swedish

Figure 3: Time series across domains for each language, with values representing the average predicted connotation over one month for the general and hate domain, and events of interest highlighted with red points.

feminist appears to be more negative than the plural *feminister* in the Swedish data. This is likely due to mentions of Feminist Initiative (FI), one of Sweden's political parties. We can see that the plural form overtakes the singular for a brief period before and after the election in September 2014. Since then, support for the FI party in Sweden has rapidly declined. This is reflected in the figure.

5.2 Event-Driven Analysis

This section explores how major socio-political events and public discourse shaped the online perception of feminists over time, across languages and domains. Fig. 3 compares the rate of predicted connotations towards feminists across domains.

Italian As expected, the general domain consistently shows higher rates of non-negative connotation than the hate domain. The hate domain maintains low levels of non-negative connotation, often below 0.25, signaling predominantly negative sentiment toward feminists. The general domain shows a gradual decline from around 0.7 (2014-2016) to around 0.6 or slightly below (2021), but the trend is relatively stable. The hate domain, in contrast, has erratic fluctuations, with many sharp spikes and drops. The general domain shows some dips, notably around 2017–2018 and 2022–2023. In 2017 we observe both a positive and negative correlation with the Donald Trump's inauguration (Jan 20, 2017), which prompted massive backlash due to his history of misogynistic remarks and policies perceived as anti-women. However, On January 21st, the day of the Women's March following Donald Trump's inauguration, Italian online discourse reflected a noticeable increase in non-negative sentiment toward feminists, driven by support and solidarity. However, critics emerged, questioning the selective outrage of feminists, with posts asking, "Where were feminists when Bush, Obama, and Clinton massacred women and children in Africa and the Middle East?". Others used dismissive or hostile language, such as calling them "femministe sinistronze radical-shit che non si indignano per la violenza degli immigrati" (leftist radical-shit feminists who don't get outraged about violence from immigrants), revealing the ideological fallacy of *whataboutism* used to discredit feminists by accusing them of a DOUBLE STANDARD, particularly in relation to religion and migration-related issues, as captured by the INTERSECTIONALITY frame.

Another key moment leading to a drop in sentiment occurred in August 2021, in response to the crisis in Afghanistan following the Taliban's return to power. As reports of Afghan women losing basic rights and freedoms spread, online discourse saw a surge of criticism for Western feminists. A recurring, sarcastic refrain—"Where are the feminists now???"—emerged, accusing feminists of being selectively vocal and absent in moments of geopolitical crisis affecting non-Western women. These comments often framed feminists as hypocritical.

In the hate domain, it was hard spotting triggering events. Throughout the negatively-connotated posts, feminists are blamed for promoting moral decay, often associated with sexual liberation. Throughout the years, there is frequent portrayal of feminists as manipulative or opportunistic, and they consider feminism to be a *lobby* that aims for the world supremacy. This resonate with the DARVO frame, which is one of the most common in the hate domain.

English Data in the general domain indicates a stable trend of low negativity remaining stable over time. This patterns persists even during discussions surrounding significant social and political events, including Hillary Clinton's nomination as the first female presidential candidate for a major party (July 28, 2016), Kamala Harris taking office as vice president (January 21, 2021), and the overturning of Roe v. Wade by the U.S. Supreme Court, resulting in abortion restrictions across numerous states (June 2022). Compared to Italian, fluctuations in the English general domain appear to be less event-driven. This may be due to the nature of the different platforms from which the data was collected. Users who write on Twitter tend to post in response to ongoing events. Reddit posts and comments, in turn, are often longer and more discursive; users propose topics of discussions rather than commenting on happenings. However, we observe that the highest level of negativity in this domain is reached in late 2018. Upon examining the period starting in September, this surge aligns with the Brett Kavanaugh Supreme Court hearings[7]. This highly publicized event catalyzed widespread discourse surrounding gender dynamics, particularly among men expressing anxiety over the potential for false accusations, with Reddit posts such as "I won't ever touch a woman anymore" blaming feminists. Such narratives, rooted in perceived male victimhood, contribute to the high frequency of the DISMISSAL OF FEMINISM frame in this domain. The English hate domain consist of comments from 2017 to 2022. The data shows consistently high levels of negativity, with occasional peaks and drops. This pattern likely reflects the nature of the forum itself, which is marked by the expression of strong opinions and highly polarized discourse. Notably, a spike in anti-feminist comments occurs in April 2018, possibly linked to the Toronto van attack carried out by Alek Minassian, a 25-year-old man who described himself as an incel (April 23, 2018). Various threads around this date refer to the attack, discussing the act itself or the impact that this may have on the perception of incels. In comments like "correct. the blood is on the hands of feminists and 'women' who created this culture. this man is nothing but a product of his environment.", or "i oppose violence but considering how the normies are reacting i really wouldn't be surprised if there aren't more attacks. [...] feminist dishonesty about what women want and their increasing hypergamy added to a pool of emotionally damaged beta males [...]", users justify the perpetrator's action and blame women and feminists and minimizing the perpetrator's responsibility. This type of content, which represents women as oppressors, privileged, and deceiving, reflects the high frequency of the DARVO frame individuated in the hate domain annotated data.

Swedish It is challenging to determine specific events which may have lead to increases in negative predictions or clear patterns given the sparsity of data for the hate domain in Swedish. Despite this, we observe a few events that appear to have been indirect catalysts for an increase in negative predictions in the Incel forums. Dr. Stefan Krakowski, a well-known scholar of Incel culture in Sweden, presented a lecture in May, 2023. This event is mentioned in one thread, after which we see an extremely low rate of predicted non-negative labels for several months. We also see a steady decline in non-negative predictions from mid 2020 to late 2021, roughly following a thread mentioning an article published in March 2020 claiming Sweden the most 'incel' country in the world.

For Reddit, we observe a dip in 2015 arround the time of a debate broadcast on Sweden's national television (SVT) featuring several members of Sweden's Feminist Initiative party. In 2017, there are several drops in non-negative predictions which correspond to specific events. The first of these is from May to September 2017 around the time of rumors for plans to hold a 'man-free' music festival following the cancellation of the popular music festival Bråvalla due to several sexual assault allegations. A smaller decrease is also observed in late 2017 to early 2018, around the time of the MeToo movement and the introduction of a bill to amended Sweden's laws on consent.

There is another drop in non-negative predictions following a court decision to convict Cissi Wallin, a controversial figure in the Swedish MeToo movement, of defamation in late 2019. Another drop is observed in March 2020 at the time of an ad campaign on Instagram for the Swedish Armed Forces seemingly targeting women. These two events appear to have had a more immediate negative response.

[7] Kavanaugh was nominated by Donald Trump as judge for Associate Justice of the Supreme Court of the United States in July 2018, but during the confirmation process he was accused of sexual assault. The accusations were made public by the *Washington Post* in September of the same year.

Overall, the average of predicted non-negative labels appears to decline in the months leading up to the 2014, 2018, and 2022 Swedish general elections. A lot of comments in these periods mention Sweden's Feminist Initiative (F!) party.

6 Conclusion

Using a combination of frame analysis and classification models, we examined large-scale online discourse to detect the connotation of *feminist(s)* across time, languages and domains. Our analysis reveals that the connotation of *feminists* is consistently more negative than its singular form over the years, except for Swedish. The main driver for this difference in Swedish is the Feminist Initiative party, which has drawn ire online since 2014. As expected, our data shows that the hate domain exhibits considerably higher levels of negativity toward feminists than the general domain in all languages. Indeed, *feminist(s)* is frequently used as a slur, often appearing in contexts associated with the INSULT frame. This contrasts with the general domain, where negative stances toward feminists are less overtly expressed, in the form of STEREOTYPES, DISMISSAL OF FEMINISM, and MISANDRY. We also found that event-driven shifts were more easily detectable in the general domain and are linked to socio-political issues. Italian displays greater linguistic inventiveness, often incorporating politically charged epithets such as *centrosocialina* (squat girl), *sinistroide* (leftoid), and *zecca* (tick/communist). In contrast, English discourse tends to represent feminists as overprivileged and power-driven, framing them in terms of their supposed higher socioeconomic status, using descriptors like *white* and *rich*. Discourse in Swedish tends to describe feminists as political manipulators, claiming that they destroy Sweden. These findings highlight that feminists are framed differently across the three languages. To further enhance the coverage and representativeness of our analysis, future work could benefit from a participatory design approach that incorporates knowledge contributed by those who are directly engaged in gender advocacy and discourse.

Limitations

This research does not come without limitations.

The results obtained using GPT and XLM-T for frame prediction were unsatisfactory, with macro F1-scores consistently falling below 0.2. While this reflects the inherent complexity of the task, performance was too low to justify detailed reporting. Future work will explore strategies to improve model effectiveness in frame prediction.

Another methodological limitation relates to our use of a binary classification model to separate general and hate domains, which achieved an macro F1-score around 0.7. Although a more accurate system would strengthen the analysis, we consider this a reasonable starting point for an exploratory study that nonetheless revealed meaningful linguistic and event-driven patterns.

The dataset also presents several coverage-related limitations. First, while our study focuses on English, Italian, and Swedish, all three are Western European languages. Expanding the analysis to include non-Western languages would be essential to develop a more globally representative understanding of feminist discourse online. Second, our keyword-based collection method successfully retrieved a wide range of relevant discourse but necessarily misses cases in which feminist identities or perspectives are evoked without the use of explicit keywords. Third, we acknowledge gaps in the temporal coverage of the dataset. Although data collection spans approximately from 2016 to 2023, the timeframes vary slightly across sub-corpora. This variation reflects the availability of data and, while not ideal, does not significantly affect the findings. Additionally, in most languages the plural term feminists is interpreted as gender-neutral; however, in Italian we did not include the masculine plural form femministi, which may have influenced observed temporal or thematic patterns. The dataset used for frame analysis is relatively small, particularly for English. This limits the generalizability of our findings. Future work will aim to expand the annotated data through additional annotators to improve coverage and reliability.

A further consideration is that the connotation of feminist can vary across regional and cultural contexts, especially in globally spoken languages like English. Since our dataset lacks geolocation information, we were unable to account for regional variation in meaning.

Finally, while we followed a structured annotation procedure, it had limitations due to external constraints such as budget and annotator availability. Systematically analyzing annotation disagreements in future work could help uncover ambiguities in how negativity is expressed.

Acknowlegments

Arianna Muti's and Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Emanuele Moscato's research was funded by the European Union - NextGenerationEU, in the framework of the FAIR - Future Artificial Intelligence Research project (FAIR PE00000013 – CUP B43C22000800006). Emanuele Moscato, Arianna Muti, and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

During part of this study, Flor Miriam Plaza-del-Arco was supported by the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), as part of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Donya Ahmadi. 2024. Between a rock and a hard place: The intersectional experiences of iranian feminists from minoritized ethno-national backgrounds. *Religions*, 15(5).

Jessica Aiston. 2024. 'vicious, vitriolic, hateful and hypocritical': the representation of feminism within the manosphere. *Critical Discourse Studies*, 21(6):703–720.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Maxime Dafaure. 2022. Memes, trolls and the manosphere: mapping the manifold expressions of antifeminism and misogyny online. *European Journal of English Studies*, 26(2):236–254.

Valerie Dickel and Giulia Evolvi. 2023. "victims of feminism": exploring networked misogyny and# metoo in the manosphere. *Feminist Media Studies*, 23(4):1392–1408.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Paolo Gajo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. 2023. On the identification and forecasting of hate speech in inceldom. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 373–384, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sara Gemelli and Gosse Minnema. 2024. Manosphrames: exploring an Italian incel community through the lens of NLP and frame semantics. In *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024*, pages 28–39, Torino, Italia. ELRA and ICCL.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Yudi Guzmán-Monteza. 2023. Assessment of an annotation method for the detection of spanish argumentative, non-argumentative, and their components. *Telematics and Informatics Reports*, 11:100068.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Ruth Lewis, Mike Rowe, and Clare Wiper. 2019. Online/offline continuities: Exploring misogyny and hate in online abuse of feminists. *Online othering: Exploring digital violence and discrimination on the Web*, pages 121–143.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Janice McCabe. 2005. What's in a label? the relationship between feminist self-identification and "feminist" attitudes among us women and men. *Gender & Society*, 19(4):480–505.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024a. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. 2024b. PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia. ELRA and ICCL.

Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right.* Zero Books, Winchester, Hampshire, UK.

Louise North. 2009. Rejecting the 'f-word' how 'feminism'and 'feminists' are understood in the newsroom. *Journalism*, 10(6):739–757.

Debora Nozza and Dirk Hovy. 2023. The state of profanity obfuscation in natural language processing scientific publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Språkbanken Text. 2024. Flashback: Livsstil.

Kalpana Srivastava, Suprakash Chaudhury, P. S. Bhat, and Samiksha Sahu. 2017. Misogyny, feminism, and sexual harassment. *Industrial Psychiatry Journal*, 26(2):111–113.

Märta Stenavi and Karin Bengston. 2020. Kvinnohat och våldshyllningar i digitala incelmiljöer. Technical report, Totalförsvarets forskningsinstitut (FOI).

Marcela Suárez Estrada, Yulissa Juarez, and CA Piña-García. 2022. Toxic social media: Affective polarization after feminist protests. *Social Media+ Society*, 8(2):20563051221098343.

Jill M Swirsky and DJ Angelone. 2014. Femi-nazis and bra burning crazies: A qualitative evaluation of contemporary beliefs about feminism. *Current Psychology*, 33:229–245.

Michael Yoder, Chloe Perry, David Brown, Kathleen Carley, and Meredith Pruden. 2023. Identity construction in a misogynist incels forum. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

A Data

Table 4 shows the data size across languages, domains, years, and sources.

B Experimental Settings

B.1 Model Training

Fine-tuning of models is performed using the Hugging Face API, with the default AdamW optimizer (Loshchilov and Hutter, 2019) and linear learning rate decay. Table 5 summarizes the training parameters used for each model. The EN and IT models has been trained on two NVIDIA RTX A5000 GPUs while the SE model has been trained on one NVIDIA A100.

B.2 Prompt

> You are a feminist annotator. Determine whether the term "feminist" (or its plural form "feminists") is used with a negative connotation in the given text.
>
> Each text will be annotated with a binary label indicating the connotation of the text:
>
> - 1 (negative) if the term feminist (or feminists) is used in a pejorative, derogatory,

Lang (Domain)	Year	# of posts	Source
IT (General)	2014	4912	TWITA
	2016	982	TWITA
	2017	7164	TWITA
	2018	22028	TWITA
	2019	39196	TWITA
	2020	45264	TWITA
	2021	38413	TWITA
	2022	29958	TWITA
	2023	2669	Pejorativity
IT (Hate)	2014	215	FdB
	2015	112	FdB
	2016	146	FdB
	2017	140	FdB
	2018	571	FdB
	2019	584	FdB
	2020	579	FdB
	2021	649	FdB
	2022	336	FdB
	2023	176	FdB
EN (General)	2014	720,019	Reddit
	2015	818,452	Reddit
	2016	760,921	Reddit
	2017	820,047	Reddit
	2018	900,150	Reddit
	2019	983,674	Reddit
	2020	948,485	Reddit
	2021	970,925	Reddit
	2022	1.025.294	Reddit
	2023	1,033,587	Reddit
EN (Hate)	2017	240	Incel.is
	2018	2104	Incel.is
	2019	1813	Incel.is
	2020	2012	Incel.is
	2021	1320	Incel.is
	2022	880	Incel.is
SE (General)	2014	5764	Reddit
	2015	6338	Reddit
	2016	4292	Reddit
	2017	3853	Reddit
	2018	2820	Reddit
	2019	925	Reddit
	2020	910	Reddit
	2021	1197	Reddit
	2022	1074	Reddit
	2023	-	Reddit
SE (Hate)	2016	23	Flashback
	2019	113	Flashback
	2020	2	Flashback
	2021	34	Flashback
	2022	10	Flashback
	2023	222	Flashback

Table 4: Stats for FEMME. In TWITA, we suspect that in 2015 our target words were not used as keywords to retrieve tweets, therefore we do not have data for that year. For 2023, the PejorativITy corpus considers only the first two months.

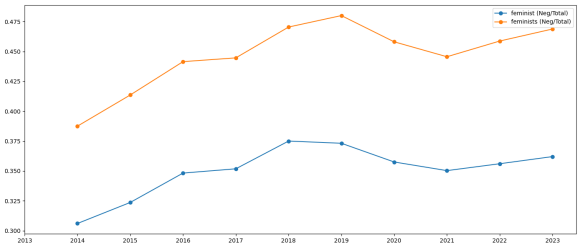
or discrediting manner, or more generally carries negative connotation.

- 0 (neutral/positive/other) if the term is used neutrally, positively, or descriptively without negative intent.

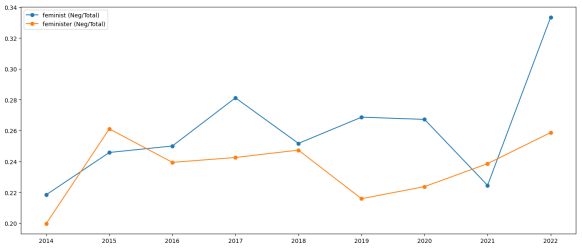
Text:

Model	Lang	Epochs	LR	Batch size
xlm-roberta	EN	10	$2 \cdot 10^{-5}$	16
alb3rt0	IT	8	10^{-5}	16
roberta-large	SE	10	10^{-5}	16

Table 5: Training parameters for model fine-tuning.



(a) English.



(b) Swedish.

Figure 4: Comparison of the ratio of negative to total mentions for feminist and feminists in English and Swedish.

C Ratio of Negative Counts of Feminist(s)

Fig. 4 shows the negative ratio for feminist and feminists in English and Swedish.

D Domain Adaptation

Table 6 shows the differences between domain-separated results. Performance in the general domain consistently outperforms the hate domain, indicating that detecting a negative connotation in more hostile or extreme contexts remains more challenging, likely due to higher lexical variation. Interestingly, Swedish shows the inverse trend with GPT-4o-mini, i.e., slightly better performance in hate (0.386) than general (0.294).

Model	Lang	Domain	F1-score
GPT-4o-mini	SE	all	0.368
		hate	0.386
		general	0.294
	IT	all	0.500
		hate	0.414
		general	0.539
	EN	all	0.920
		hate	0.596
		general	0.799
AlBERTo	IT	all	0.700
		hate	0.675
		general	0.657
XLM-T	EN	all	0.920
		hate	0.866
		general	0.951
AI-Sweden RoBERTa	SE	all	0.682
		hate	0.564
		general	0.730

Table 6: Macro F1-scores for GPT-4o-mini and fine-tuned models across different domains.