Debate around feminism in Telegram. Analysing affective and structural polarisation.

Author: Kravchenko Diana (1001307)

Introduction

The problem of attitude towards feminism in Russian society is a significant one because the position of the movement is very unstable. The country continues to be patriarchal in many respects. For instance, the share of women in the Federal Assembly in 2023 was 16.4%, which in the world ranking put Russia in 138th place between Libya and Belize.[1] Furthermore, the prevailing number of initiatives designed to strengthen the role of women in Russian society are closed. At the same time the words “feminism” and “feminist” in public discourse acquire not only negative connotations but are also used to express contempt or ridicule.[2] Importance of a woman as an independent political actor is significantly reduced, the focus shifts to her role as a means to implement the state's demographic policy.

Therefore, the topic of attitude towards feminism in Russia needs to be studied to determine whether the political discourse reflects the real opinion of the people. The work aims to determine if there is affective polarisation of feminist and antifeminist communities, as well as to evaluate the structure of the divide, if there is structural polarisation.

To answer the research question, social network analysis and sentiment analysis will be conducted on data from Telegram.

Theoretical framework

Term **polarization** is used to denote a process of division of society on a number of relatively big groups based on their radical differences in opinions, values or attitudes towards a specific social or political phenomena[3]. More and more research is focusing on online platforms as places where polarised views are demonstrated. They are viewed as mechanisms that enhance polarisation. Due to the intensity of their influence, social media is the most notable form of online platforms.

[1] Inter-Parliamentary Union (IPU). (2023).

[2] Belyaeva, G. F. (2008).

[3] Fiorina, M. P., & Abrams, S. J. (2008).

The studies outline two major types of polarization: structural polarisation and affective polarisation. The term **structural polarisation** is used to describe a situation where social and physical space become divided in accordance with political stands and values of individuals[4]. There are a few ways of manifestation of this phenomenon, one of which is *selective exposure* - avoidance by an individual of unpleasant and undesired topics[5]. Another manifestation is the existence of *echo-chambers* - isolated communities, information bubbles where interactions occur between like minded individuals[6] - homophilic interactions[7]. While the term **affective polarisation** denotes a state in which individuals "love" their we-group and treat with animosity their they-group[8].

Social media represents "Internet channels that allow users to opportunistically interact and selectively represent themselves, either in real time or asynchronously, with both a wide and narrow audience that benefits from user content and the perception of interaction with others."[9]. The structures that dictate the interactions within the platform - **the frames**[10], can influence the degree of polarisation or of the toxicity of the platform. Nevertheless, it cannot be denied that all social media platforms have a number of common features. First of all, they provide an opportunity for mass communication and interaction with a wide anonymous audience. For a long time it was believed that this would allow to form a free space for discussions on social media[11]. However, this turned out to be difficult due to two reasons: the instability of social norms and the general tendency to **homophilic interaction**. The tendency to homophily is manifested in the fact that social media communities are organized around common topics of interest, and it is in these communities that users find new "friends"[12]. Moreover, the very structure of social media promotes homophily, as it is often built on the basis of recommendation systems offering users information based on their previous searches and interests.

[4] Banks, A., Lenz, G. S., & Wagner, M. W. (2021).
[5] Ibid.
[6] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015)
[7] Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021).
[8] Iyengar, S., & Westwood, S. J. (2015).
[9] Carr, C. T., & Hayes, R. A. (2015).
[10] Yang, A. (2015).
[11] Khalili, B. G., Quraishi, T., & Fazil, S. (2024).
[12] Carr, C. T., & Hayes, R. A. (2015).

Methodology

Data

The corpus of analysed texts consists of the posts and comments to them from 5 feminist and antifeminist Telegram groups collected via Telegram API for the time period of 01.01.2024 - 01.01.2025. The groups were determined based on their popularity. Overall, for the given period 7829 posts were collected, 4322 from antifeminist groups and 3507 from feminist ones. There were 32828 comments from profeminist groups and 163677 comments from anifeminist groups.

In the end each post was represented by the text of the post, a list of all comments, containing textual information and ids of the users who wrote them, as well as id of the channel from which the post was collected and ids of the channels which were mentioned or from which the post was forwarded to the group.

Analysis

Sentiment analysis

On the next stage for each post two parameters were calculated: score of polarisation and score of expressed sentiment. In order to obtain these two parameters, a NLP model DeepPavlov was utilised. It uses RuBERT and was trained on the RuSentiment dataset which consists of data collected from Vkontakte and was specifically designed for analysis of relatively short Russian texts in informal settings. The output of the model consists of two marks: label and score. Label can take three values: positive, negative and neutral, while score signifies the intensity of assigned sentiment. Therefore, each comment obtained a label and a score.

However, the model does not analyse texts longer than 512 tokens. Hence, longer comments were split in chunks and the parameters were obtained for each chunk. In order to obtain an **overall score for the comment**, the code computes overall positive, negative and neutral scores by summing up scores for each chunk. Then the label is assigned based on which of the overall scores is the biggest. Then the score for the comment, if for example its label is neutral, is calculated with the following formula:

$$sentiment_{comment} = \frac{neutral_{chunk}}{n_{neutral}} - \frac{negativel_{chunk}}{n_{negative}} - \frac{positivel_{chunk}}{n_{positive}}$$

where $sentiment_{comment}$ - the sentiment score for the comment, $label_{chunk}$ - sum of sentiment scores with the respective label for each chunk, n_{label} - is the number of chunks with neutral, negative and positive label, respectively.

Calculation of the **polarity score** will be based on the formula from Dang-Xuan, L., Stieglitz, S., Wladarsch, J., & Neuberger, C. (2017). In the former work the following formula is proposed:

$$emotion_{auto} = positive - negative - 2 ,$$

where $emotion_{auto}$- the rate of polarisation, $positive, negative$ - sentiment scores of the publications, where positive can obtain values from 0 to 1 and negative from -1 to zero.

While the formula allows to capture emotional sentiment of the publication, it may lead to exaggeration of real level of polarisation since neutral comments are completely excluded from the analysis. Therefore, the modified formula for **polarisation score** for the post look the following way:

$$pol_{post} = 1 - \frac{neut_{post}}{n_{post}} - \left|\frac{negative_{post} - positive_{post}}{n_{post}}\right|$$

where pol_{post} - polarisation score for the post, $positive_{post}, negative_{post}, neutral_{post} -$ overall scores for respective labels computed as sums of the score of comments for a given post.

Values of $pol_{post} \in (0; 1)$, where $|pol_{post}| = 1$ signifies polarisation..

As for the **score of expressed sentiment** it is used to determine what was the predominant sentiment expressed in the comment section. It does not include neutral comments because its main purpose is to indicate general attitude towards topic and not the degree of the divide.

$$sentiment_{post} = \frac{\sum_{i=o}^{k} positive_{post_i} - \sum_{i=o}^{l} negative_{post_i}}{k+l},$$

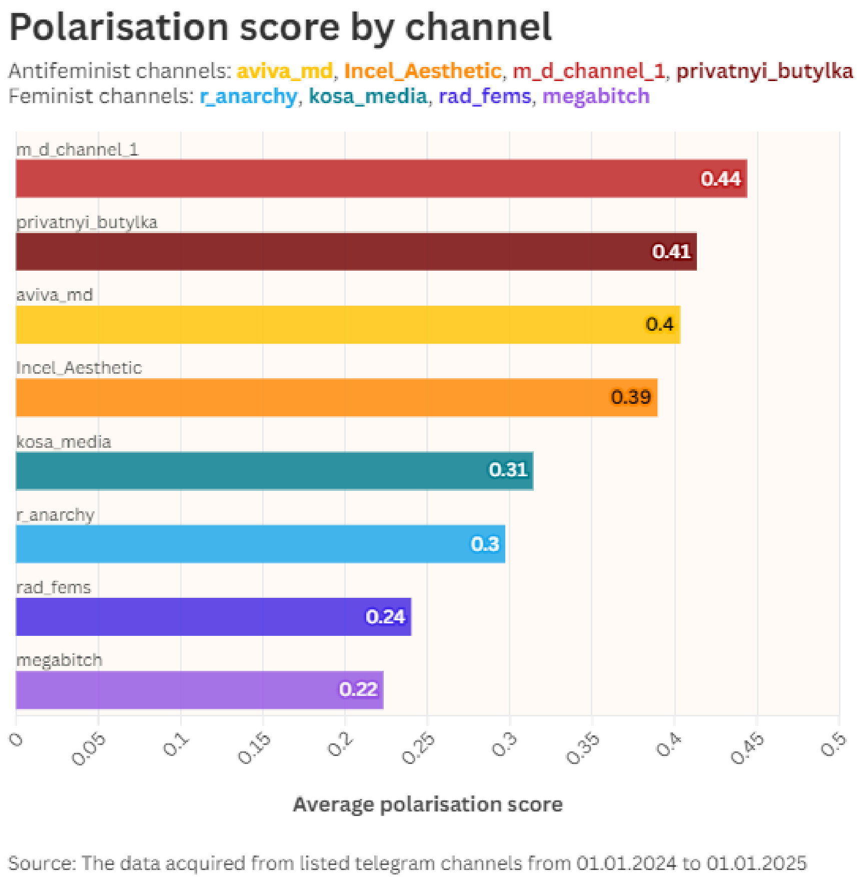
where k - number of posts with positive overall sentiment, l - with negative. $positive_{post}, negative_{post} -$ overall scores for respective labels computed as sums of the scores of comments for a given post.

$sentiment_{post} \in [-1, 1]$, where $sentiment_{post}$ = -1 means that all of the comments were labelled as *negative* with the score = 1 and vice versa for $sentiment_{post}$ = 1.

Results

Sentiment Analysis

Analysis of polarisation score computed based on sentiments shows that the groups are not affectively polarised. The average polarisation score equals to 0.33. In combination with a mean sentiment score equal to -0.36 that would indicate that discussions in the researched channels are mostly neutral with a slightly negative sentiment, which is not surprising for online communities. However, the examination of average scores by channel shows that for antifeminist channels the score is 33% higher on average. Therefore, it is necessary to evaluate two groups separately.


Polarisation score by channel

Source: The data acquired from listed telegram channels from 01.01.2024 to 01.01.2025

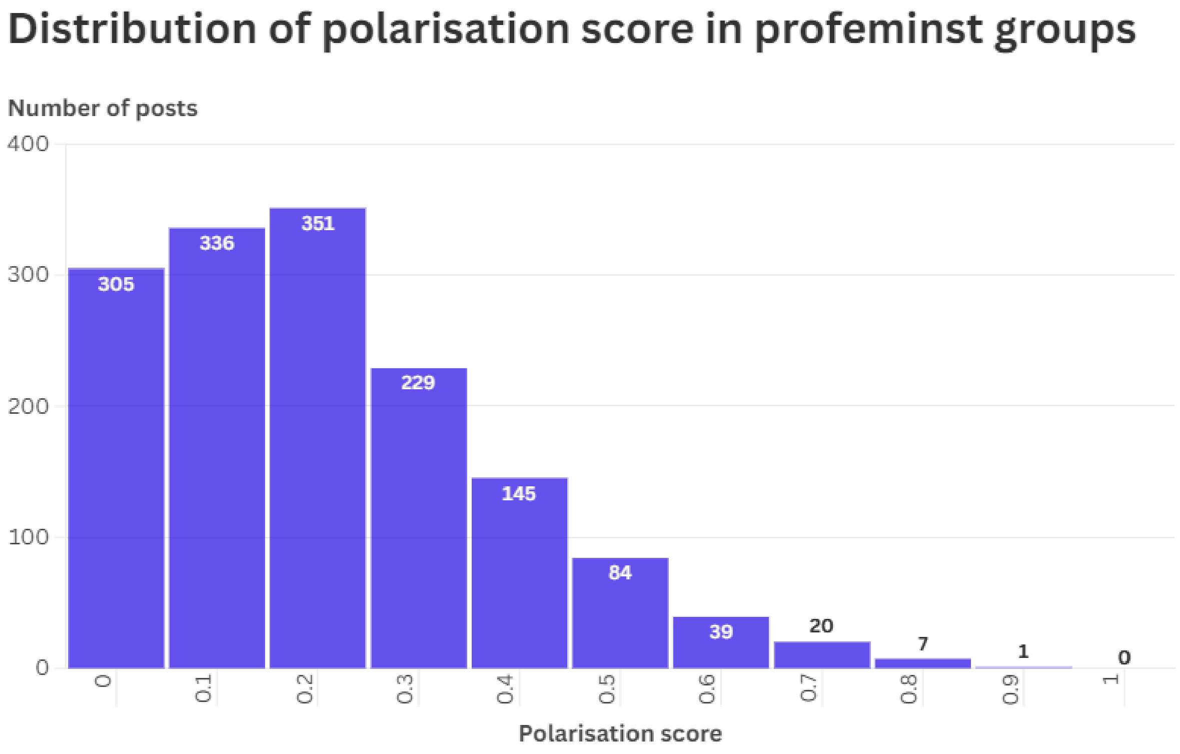
Figure 1. Average polarisation score by channel

As for the profeminist groups, the distribution of polarisation score is presented by the figure 2. The mean score is 0.26, maximum score equals to 0.97, and the share of polarised posts ($pol_{post} > 0.7$) is only 2%.. There are no signs of polarisation within profeminist communities. This can be attributed to the fact that all of them have strict rules for the behaviour within the discussion channels, users are banned for hate speech and moderators delete highly toxic comments. Moreover, all of them are under 50000 subscribers which allows for a more thorough control. As for the most “polarised” channel : kosa-media, it has the biggest number of followers out of all the channels and focuses mostly on “news about women”, thus it might attract viewers from completely different backgrounds, while other channels’ content covers theoretical questions and attracts a specialised audience. The distribution of sentiment score further proves the thesis of absence of affective polarisation, as it is centered around 0 with the mean value of -0.1.

For antifeminist comments mean polarisation score is not much higher, it equals to 0.4, nevertheless, the share of highly polarised discussions equals to 2%, the same as for feminist ones. The highest score is taken by *m_d_channel_1* In contrast to the feminist community, it has the lowest number of followers. The significant difference between two communities is in sentiment, for antifeminist channels it is 6 times more negative and equals to -0.61, while the maximum score is 0.09. Based on the sentiment score, one may argue that researched antifeminist communities have a more toxic environment.

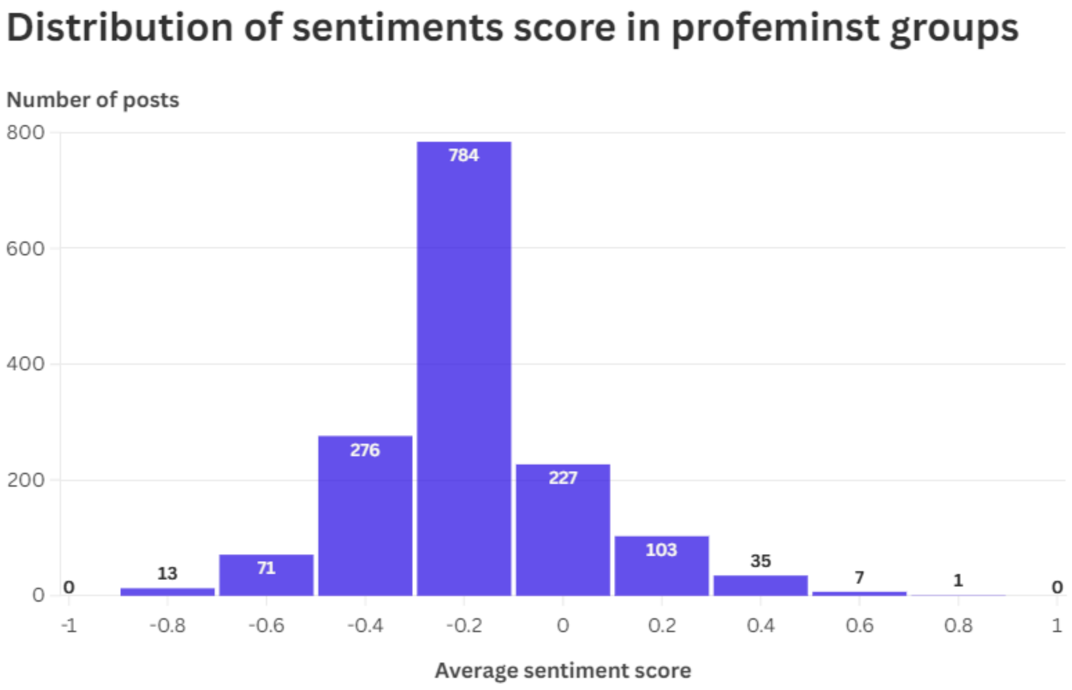
The following conclusions can be derived from the sentiment analysis: firstly, there is no indication of existence of affective polarisation between the two communities since polarisation scores are relatively low. Secondly, analysed antifeminist communist are more toxic than feminist ones.

Nevertheless, in many respects the demonstrated lack of affective polarisation and absence of animosity can be explained by structural polarisation - the researched groups exist within echo-chambers, interacting only with like-minded individuals, thus, they express the same sentiment, To examine this thesis, a network of channels and a network of users was constructed.


Distribution of polarisation score in profeminst groups

Average polarisation score: 0.25
Share or polarised discussions (score > 0.7) = 2%

Figure 2. Distribution of polarisation score in profeminst groups


Distribution of sentiments score in profeminst groups

From -1 (negative) to 1 (positive).
Average sentiment score = -0.1.

Figure 3. Distribution of sentiment score in profeminst groups

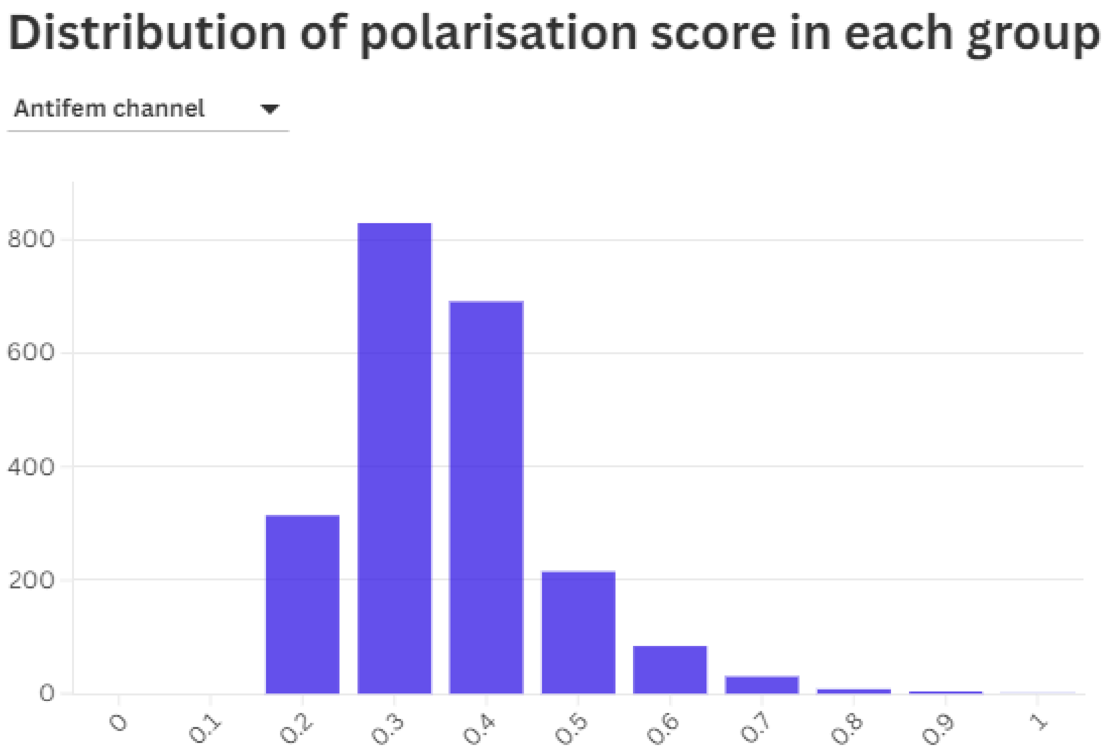


Figure 4. Distribution of polarisation score in profeminst groups

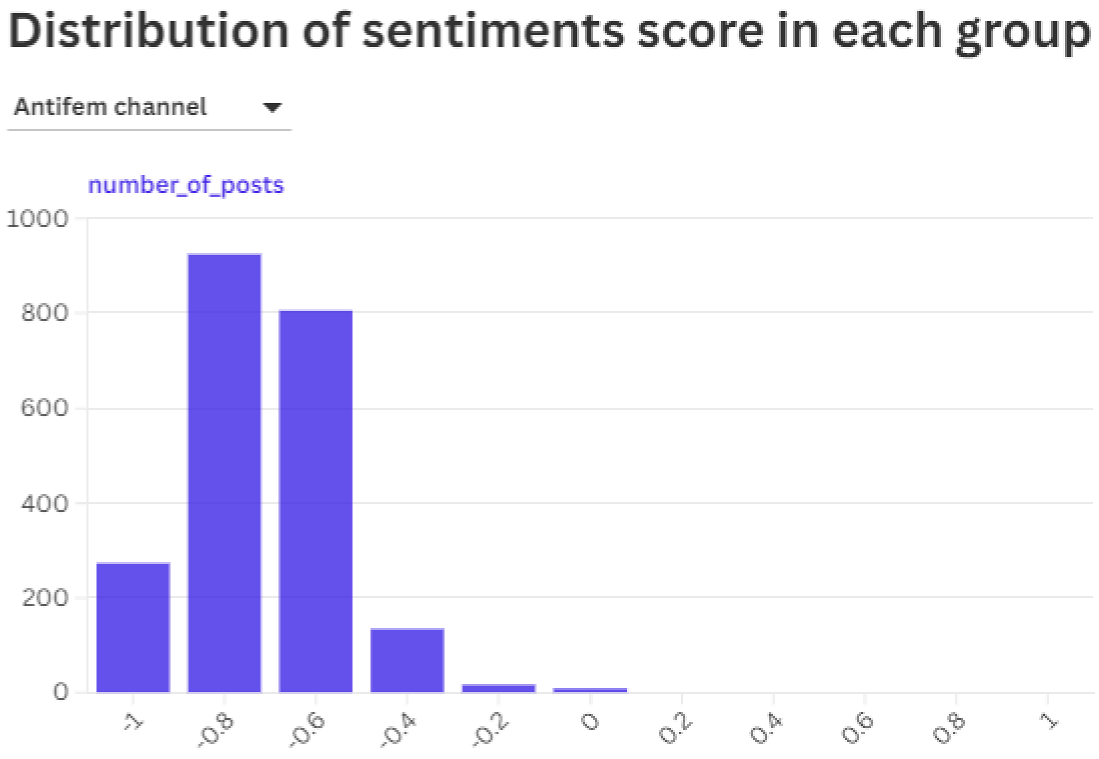


Figure 5. Distribution of sentiment score in antifeminst groups

Social Network Analysis

Channel's Network

The network construction was based on the following algorithm:

1) For each channel mentioned in the post or for the channel from which the post was forwarded a list of main_channels (the channels from which the information was gathered) was created . If there was only one channel in the list, then the color of the node is the same as the color assigned to the main channel. However, if multiple main channels forwarded information from the channel, it was assigned the label 'multiple'.
2) For each channel a number of overall mentions was calculated which is reflected on the graph using the size of the nides.
3) For each channel the number of mentions by the main channel was calculated which is reflected using the size of the edge.

On the graph (Figure 5) we can see three connected components: a network of two antifeminist channels, *r_anarchy* channel network, and a network of feminist channels with one common node with an antifeminist channel *Incel_Aesthetic.*

The form of network shows the existence of structural polarisation between feminist and antifeminist communities. The following articulation points were found: *rad_fems, kosa_media, megabitch, femagainstwar, MFRradfem, ekaterina_mizulina, Incel_Aesthetic, aviva_md, privatnyi_butylka, r_anarchy*. The only one which is not a main channel is *ekaterina_mizulina*, it is the one connecting feminist and antifeminist networks, which can be gathered from the graph(figure 5).

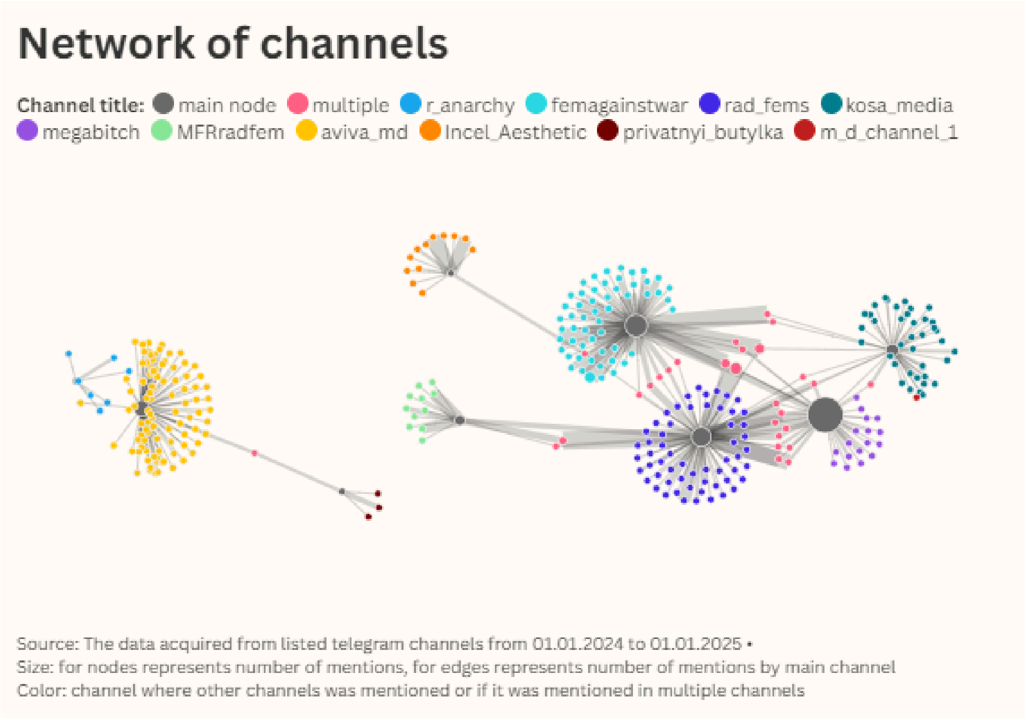
As for the antifeminist network it is divided into two subcomponents: *Incel_Aesthetic (*channel with one articulation point to the feminist network*)* and *aviva_md* and *privatnyi_butylka* cl. The latter is connected by one articulation point. Moreover, one of the channels is not on the graph at all (*m_d_channel_1*). Hence, antifeminist channels themselves exist with echo-chambers in terms of gathered information and tend to produce original content and not forward in from third parties.

The table below provides statistical evidence for the conclusions drawn from the graph as it calculates co-participation of different information channels. On average main channels have 1.167 commonly mentioned channels which is due to the input of feminist networks. There on average, the main channels mention 2.7 of the same channels.

Average Foci Overlap:	1.167
Average Foci Overlap Feminist:	2.733
Average Foci Overlap Antifeminist:	0.333
Average Foci Overlap Cross Community:	0.056

Table 1. Average foci overlaps for different components of the channel's graph

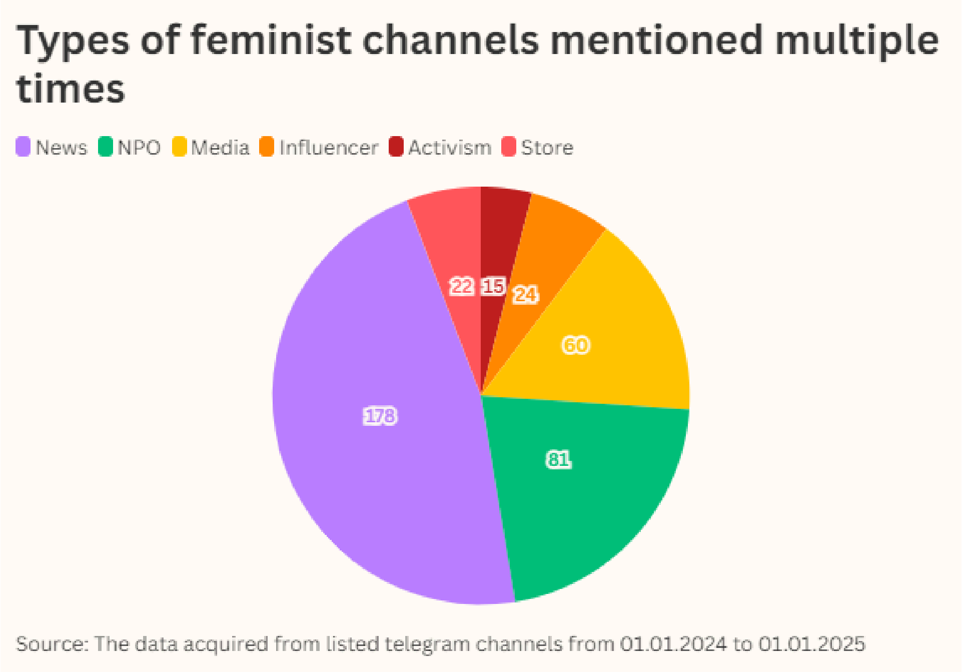
As for the overlapping nodes, oppositional news channels or news channels focusing on feminist discourse account for 46% of all mentioned (Figure 6), with the most mentions to *glasnaya_media* . *Glasnaya_media* channels focuses on women's perspective and also engages into activist activity. The second biggest share is taken by Non-profit organisations, where *ovdinfo* mentions constitute 55% of overall mentions. *Ovdinfo* is an organisation which helps people seek juridical help, for instance advocates, specifically focusing on cases of political repression.


Network of channels

Source: The data acquired from listed telegram channels from 01.01.2024 to 01.01.2025 •
Size: for nodes represents number of mentions, for edges represents number of mentions by main channel
Color: channel where other channels was mentioned or if it was mentioned in multiple channels

Figure 5. Network of channels[13]

[13] The graph is interactive and can be accessed via this link


Types of feminist channels mentioned multiple times

Source: The data acquired from listed telegram channels from 01.01.2024 to 01.01.2025

Figure 6. Types of feminist channels mentioned multiple times[14].

User's Network

User's network can be considered an **affiliation network**. As the connections between individuals are defined in terms of their affiliation with one of the main channels.

Overall, there are 11311 unique users that produced 196758 comments. 62% users are from feminist groups and 38% from antifeminist ones and 0.04% users engaged into discussion in both types of main channels (Figure 7), while 2.5% engaged into discussion in multiple channels in general. This finding can be explained by the fact that the scope of researched channels is relatively low, however, computational resources possessed by the author's devices did not allow inclusion of more users.

Analysing graph as bipartite allowed to calculate how many users on average are engaged into the discussion with multiple channels. Table 2 shows that feminist network continues to be the most interconnected, while on average the two researched communities have less than 1 common user.

However, the value of Average Foci Overlap Cross Community signifies that there is a connected component between the two communities and the following nain channels are included in it: *rad_fems, m_d_channel_1, kosa_media, r_anarchy, privatnyi_butylka, Incel_Aesthetic, aviva_md, megabitch.* Cross-community the most common users share

[14] The graph is interactive and can be accessed via this link

rad_fems and *privatnyi_butylka* with 5 users commenting in both channels[15], overall the most users share *rad_fems* and *kosa_media* with 62.

Average Foci Overlap:	4.386
Average Foci Overlap Feminist:	8.667
Average Foci Overlap Antifeminist:	5.167
Average Foci Overlap Cross Community:	0.583

Table 2. Average foci overlaps for different components of the user's graph

Therefore, analysis of the user's graph shows that while users are not fully divided along community lines, their interactions are predominantly homophilic. Hence, the researched communities are structurally polarised.

Conclusion

Sentiment analysis has shown that there is no affective polarisation within the researched channels as polarisation scores are relatively low for both types of channels. However, the sentiment scores differ significantly, the one for antifeminist channels benign six times lower with a value of -0.61. The overall predominance of negative sentiment is not surprising for online discussions, however, it is shown that the antifeminist discourse is more toxic.

Analysis of networks allowed to establish the existence of structural polarisation within the researched communities as both channels and users exist within echo-chambers. Though, the most surprising is that users within the communities are also not very well connected as well with only 2,5% engaged into multiple channel's discussions. .

[15] Other cross-community common users are shared by ('rad_fems', 'Incel_Aesthetic'): 2, ('kosa_media', 'Incel_Aesthetic'): 1, ('kosa_media', 'privatnyi_butylka'): 1, ('kosa_media', 'm_d_channel_1'): 1, ('megabitch', 'Incel_Aesthetic'): 3, ('megabitch', 'privatnyi_butylka'): 1.

References:

Belyaeva, G. F. (2008). Political activity of women in Russia. *Questions of State and Municipal Administration*, (1), 143–164.
(in Russian: Беляева, Г. Ф. Политическая активность женщин в России. Вопросы государственного и муниципального управления, (1), 143–164).

Inter-Parliamentary Union (IPU). (2023). Women in power in 2023: New data shows progress, wide regional gaps. Retrieved from https://www.ipu.org/news/press-releases/2023-03/women-in-power-in-2023-new-data-shows-progress-wide-regional-gaps

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, 11, 563–588.

Banks, A., Lenz, G. S., & Wagner, M. W. (2021). #polarizedfeeds: Three experiments on polarization, framing, and social media. *The International Journal of Press/Politics, 26*(3), 609–634. https://doi.org/10.1177/19401612211014887

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science, 26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, 23(1), 46–65.

Stieglitz, S., Wladarsch, J., & Neuberger, C. (2017). An investigation of influentials and the role of sentiment in political communication on Twitter during election periods. In *Social Media and Election Campaigns* (pp. 168-198). Routledge.

Kligler-Vilenchik, N., Baden, C., & Yarchi, M. (2020). Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and WhatsApp. *Social Media + Society*, 6(3), 2056305120944393.

Khalili, B. G., Quraishi, T., & Fazil, S. (2024). The influence of social media on human and social communications: A sociological study. *Journal of Social and Humanities*, 2(1), 40–48.

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1–2), 98–139.

Yang, A. (2015). Building a cognitive-sociological model of stereotypes: Stereotypical frames, social distance, and framing effects. *Howard Journal of Communications*, 26(3), 254–274.