Ethnic minorities in Russia on Telegram:

Clusters of concerns, 2022-2024

PARTICIPANTS

Artom Banissi
Haohan (Lily) Hu
Richard Rogers
Mila Rossokhatska
Gulnaz Sibgatullina
Riccardo Ventura

INTRODUCTION

Russia's large-scale invasion of Ukraine in 2022 and the subsequent crackdown on oppositional media and activists in Russia have turned **Telegram** into a powerful platform for expressing dissent and discussing anti-state perspectives. Notably, there has been significant growth in **channels representing ethnic minority groups in Russia** advocating for increased regional autonomy, ranging from systemic federation reforms to complete political independence from Moscow. Despite the importance of these developments, the cluster examination of such Telegram channels, analysis of discursive framings, and the evolution of this ecosystem over the last two and a half years has received only limited attention. Drawing on the set of telegram channels related to different ethnic minority groups in Russia, we zoom on the ecosystem of oppositional ethnic minority channels.

MAIN RESEARCH QUESTIONS

1. How many clusters of channels focused on Russia's ethnic minority issues can be distinguished as of July 2024, and how these clusters can be best characterised?

2. What are the main themes discussed in these clusters and are there overlapping themes across clusters?

Fig. 1

Project 1: Clustering

Methodology

Data collection

We started by determining seed Telegram channels, which were then grouped into three categories: channels representing ethnic minorities in the **North Caucasus** (12 channels, Region 1); the **Volga-Ural region** (11 channels; Region 2); **the North, Siberia, and Far East** (6 channels; Region 3) (for the location of the regions on the map of Russia, see Figure 1). Additionally, we created a group of channels that claimed to represent all ethnic minorities in Russia and advocated, to varying degrees, for the **defederalization** of the country (9 channels; Defederalisation group). For each group, we launched a 4Cat crawler (Peeters & Hagen, 2022) at depth 1 to discover new channels that served as sources of forwarded messages posted in the original seed channels.

Data curation

The data collected for Regions 1-3 was merged and organized into a bipartite network. In this network, the nodes represent Telegram channels, with the number of forwards as an attribute. The graph was then analyzed using Gephi (Bastian, Heymann, Jacomy, 2009). We performed community detection using the modularity function with a resolution of 0.9 and detected 10 clusters (see Figure 2)

Finding: Ethnic nationalism

The Volga-Ural region stands out as the most prominent across clusters. Within this region, there is a differentiation among ethnic groups, with distinct (sub-)clusters for the Chuvash, Tatars, Bashkirs, and Nogais. Surprisingly, the Caucasus region is less represented in the clusters. This could be attributed to several factors: a) frequent deletion of channels: 3 out of the 12 original seed channels from the Caucasus are unavailable as of July 11; b) dominance of Islam-related channels or those focused on regional news, which were not included in the original seed, potentially skewing the representation.

Finding: Cross-region communication

The clusters reveal that different regions are not isolated but share a significant amount of content across regions. Clusters 1, 4, 6, and 9 show active cross-region communication, especially between the Caucasus and the Volga-Ural regions. There is less, but still notable, communication between the Volga-Ural region and the North. Siberia remains largely disconnected from other regions, indicating limited interaction or shared content with the Caucasus, Volga-Ural, and Northern regions.

Clusters network

Cluster: 0 (less than 1%)
- Characteristics: detached from the rest, a cluster around @ar_mukhametov.
- Region: Tatarstan (Volga-Ural region); but more focused on the shared Muslim identity

Cluster 1: (about 16 %)
- Characteristics: a merger of relatively marginal channels, with openly pro-independence agendas.
- Region: Caucasus, also Chuvashia (Volga-Ural)

Cluster 2: (about 6%)
- Characteristics: a merger of politically proactive, critical voices (e.g., @neyasin) speaking about issues in the region.
- Region: Mostly Tatarstan, also Bashkortostan (both Volga-Ural)

Cluster 3: (about 8%)
- Characteristics: a merger of two channel ecologies that advocate for independent republics.
- Region: Mostly Tatarstan, also "Nogaistan" in Astrakhan region (both Volga-Ural)

Cluster 4: (about 13,5%)
- Characteristics: a merger of several channels, speaking critically of the issues in the region and the state authorities
- Region: Mostly Caucasus, also Bashkortostan

Cluster 5: (about 8%)
- Characteristics: a cluster of Tuva- and Buryatia-related channels
- Region: Siberia and the Far East

Cluster 6: (about 7,5%)
- Characteristics: a group of channels representing small ethnic minority groups
- Region: Volga-Ural (Chuvashia) and the North (Finno-Ugric communities)

Cluster 7: (about 20%)
- Characteristics: a cluster around @The_Circassian_Times channel
- Region: unrecognised Circassia (Caucasus)

Cluster 8: (about 8%)
- Characteristics: a cluster around @tpolit
- Region: focus on Tatarstan (Volga-Ural), but tackles broader issues

Cluster 9: (about 13%)
- Characteristics: a cluster around @uralistica_com, focusing on Finno-Ugric peoples
- Region: the North, also Volga-Ural

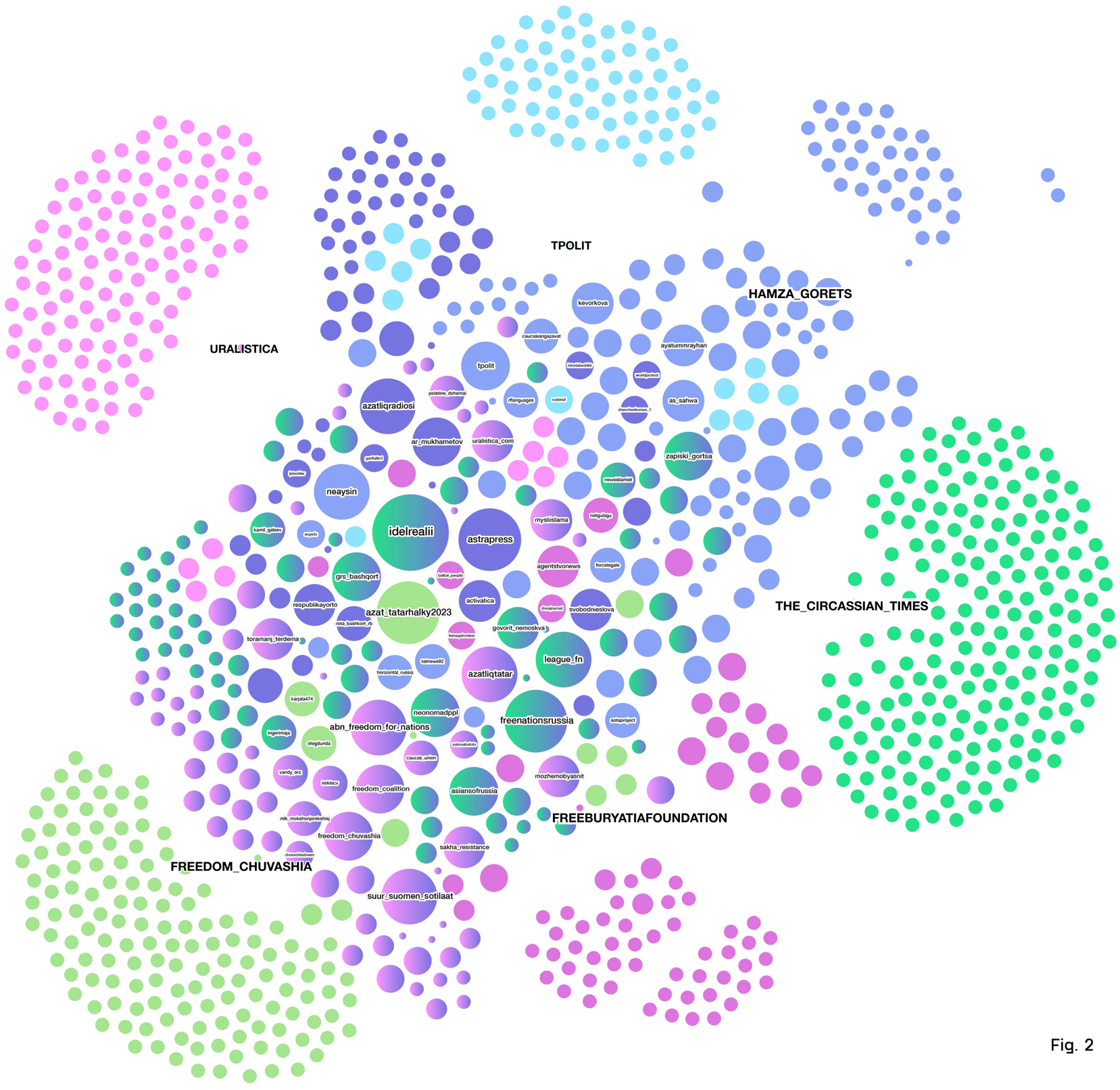


Fig. 2

Project 2: Narrative Analysis

The project focused on the analysis of the narratives present in each of the 10 clusters identified in Project 1. We decided to test a Large Language Model (LLM) on their capacity to automatically detect **narratives in each cluster**, and eventually **overarching themes.**

Methodology

1. The first step was to ask ChatGPT (the only LLM that could handle complex analysis of a corpus in Russian) to determine up to five narratives per Telegram post in each cluster. Given the multi-linguistic nature of our data (mainly Russian, but also Tatar, Finnish, Estonian, etc.), we experimented with the language of our prompt. We focused on two well-used languages: English and Russian. The analysis in English was richer and more extensive compared to the analysis in Russian, which tended to be more generalized; therefore, our final used prompt was in English (Fig. 3). We designed our prompt based on the best practice from Törnberg (2024). To stimulate an effective prompt for ChatGPT, we edited the prompt with the assistance of PromptPerfect to improve the accuracy of the generated results. Finally, we used Prompts Compass (Borra, 2023), a tool providing access to local LLMs and platform APIs, to process our textual data in CSV files using GPT-3.5. Prompts Compass treated each line of the input CSV file as a separate input.
2. In the second step, we aimed to identify up to seven overarching themes for all detected narratives per cluster. In **LLM** (ChatGPT), we encountered a recurring issue: there was a bias towards the first five narratives when analyzing data per cluster, and a bias of including data from only the first five clusters out of ten when analyzing all clusters together. This issue aligns with findings from existing studies, which indicate that LLMs are sensitive to the wording and order of prompts (Pezeshkpour, Hruschka, 2023).
3. We, therefore, decided to continue through two different routes: the quantitative way of employing **NLP** technique on the corpus, and the qualitative way of **close-reading** the narratives identified in step 1. For the **NLP**, we employed BERTopic, a robust topic modelling technique, to analyze narratives and identify prevalent themes. We preprocessed the text by removing non-alphabetic characters and stopwords using the nltk library, followed by tokenization and lemmatization to normalize words. From the lemmatized tokens, we generated bigrams (see Figure 5). We then used the BERT model to create embeddings for each narrative and applied HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) to cluster these embeddings into distinct topics. This process yielded up to 15 topics, each characterized by its most representative words and analyzed for thematic coherence (see Figure 4). However, the results were generally too broad to be useful for the purposes of the project.
4. For the **qualitative analysis**, we repeated the data cleaning process mentioned above. For the generated bigrams, we counted the frequency of each bigram to identify prevalent word pairs, thus highlighting key themes within the dataset. For each community, we created lists with the highest occurring (>10) bigrams, which were then grouped into overarching themes. For the results per cluster, see Figure 6.

Fig. 3

```
You are a narratologist tasked with identifying five narratives within a
provided Russian text. Note that a narrative is not merely describing
the events in the text, but uncovering the subtle and implicit
messages that the author seeks to express. Narratives should be
connected to an overarching set of aims or values.

Please follow these steps:

- Analyze the provided text.
- Identify up to five narratives, each summarizing the subtle and
implicit messages in 10-15 words.
- For each narrative, add an overarching theme into round brackets
in Russian. If none, write NA inside the round brackets.
- If fewer than five narratives are found, list as many as you can.
- If no narratives are found, write NA.

Output the result in the following format:
Narrative Summaries in Russian.

The text to analyze is provided below:
{user_input}
```

Fig. 4

52 национальный идентичность
46 борьба свобода
45 национальный движение
41 свобода независимость
35 культурный наследие
32 борьба независимость
31 татарский народ
30 коренной народ
27 поддержка национальный
27 протест против
23 независимость татарстан
19 российский империя
18 активный участие
18 независимый государство
17 свобода слово

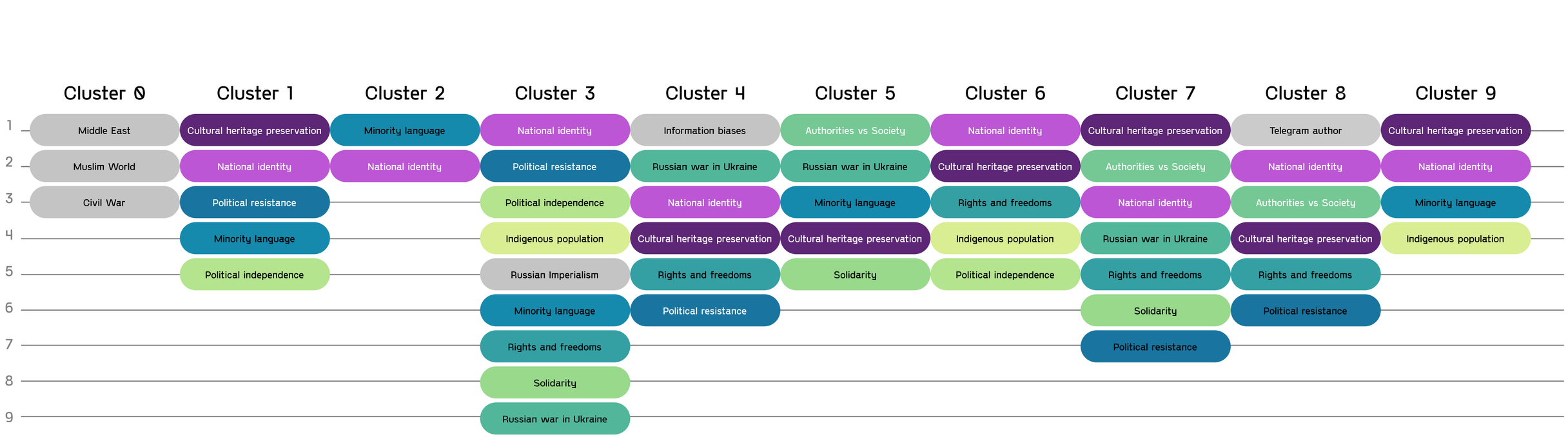
Fig. 5

Finding: Center and ethnic minority regions

The majority of the detected overarching themes are shared across all clusters, with the exception of cluster 0, an outlier identified in Project 1. The themes concerning ethnic minority communities are, rather expectedly, related to the preservation of minority cultural heritage and language, critique of the federal center, and forms of possible political resistance.

Finding: Anti-imperialism

Russian war in Ukraine is a major theme in four out of ten clusters. The decolonial framing, including critique of Russia as an empire, is clearly detectable only in cluster 3, though cluster 4 contains a significant number of narratives focused on revealing instances of racial discrimination, specifically against Asiatic peoples in Russia.



General discussion

Methods

Working with Cyrillic (Russian language, in particular) **remains a challenge for Large Language Models** that are English-language focused (or French for Mistral). These models may not fully capture the cultural nuances and regional variation, potentially leading to misinterpretation or loss of context. Existing biases in large language models (LLMs) persist when applied to Russian language data, making results often unreliable. Nevertheless, it should be recognised that when working with **multilingual datasets**, ChatGPT has managed to correctly identify most of the minority languages, specific for the Russian context, and provide generic analysis.

The nature of Telegram channels is highly dynamic. Given the volatility of channel availability (channels being deleted), data collected at different times may vary significantly, affecting longitudinal studies and the consistency of findings.

Data related to ethnic minority communities is often heterogeneous in terms of language. Various dialects, mixed-language content, and code-switching within messages add layers of complexity to the analysis. **The types of content** shared in these Telegram channels are **diverse**, including photos, videos, and text. This heterogeneity requires robust multi-modal analysis techniques to accurately interpret and integrate different forms of data.

Findings

Russia's brutal attack on Ukraine and refusal to recognise its sovereignty has understandably sent shock waves also through ethnic minority groups in Russia. The identified clusters and the overarching themes demonstrate that many of the concerns are not region specific, but shared. The overarching themes related to **ethnic minority identification** is a thread that unites all communities. There is a continiously strong attachment towards national history, cultural heritage, religious believes, which is in line with the main academic debates on ethnic minority nationalism in Russia (e.g., Prina, 2015).

At the same time, the **grievances** related to what can be referred to as the "ontological security" of these ethnic minorities, already strongly present before 2022 (Giuliano, 2011), **have found renewed expression** after the full-scale invasion, particularly **on Telegram** – the media platform convenient for group discussions, that is believed to provide anonymity and secure communication channels. These grievances are concerned with the protection and preservation of elements considered as defining an ethnic minority, such as language, cultural practices, and oral historical traditions. The data shows that **the context of war has intensified ethnic minorities' search for allies**—communities with similar political goals both inside and outside of Russia. This is evident in the types of detected clusters, the overarching theme of solidarity, and connections to other political actors, such as the Russian Volunteers' Corps fighting on the side of Ukraine (Raspad, 2023). **The driving element** behind the struggle for the protection and preservation of ethnic minority identity, as well as the search for possible political allies, **is the critique of the Russian federal center.** Despite variations across ethnic minorities and differing relations of minority regions with the Kremlin, Telegram channels that speak on behalf of these minorities are explicit in their criticism of Moscow's actions, both domestic and foreign.

Acknowledgements

We thank Stijn Peeters and Erik Borra for their assistance.

Bibliography

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
Borra. E. (2023). ErikBorra/PromptCompass: v0.4 (v0.4). Zenodo. https://doi.org/10.5281/zenodo.10252681
Giuliano, E. (2011). Constructing grievance: Ethnic nationalism in Russia's republics. Cornell University Press.
Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. Computational Communication Research, 4(2), 571–589.
Pezeshkpour, P., & Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. arXiv preprint arXiv:2308.11483.
Prina, F. (2015). National minorities in Putin's Russia: Diversity and Assimilation. Routledge.
Rogers, R., & Zhang, X. (2024). The Russia–Ukraine War in Chinese Social Media: LLM Analysis Yields a Bias Toward Neutrality. Social Media+ Society, 10(2), 20563051241254379.
Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. arXiv preprint arXiv:2402.05129.
Pezeshkpour, P., & Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. arXiv preprint arXiv:2308.11483.