Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers

Anonymous submission

Abstract

Othering, the act of portraying outgroups as fundamentally different from the ingroup, often escalates into framing them as existential threats—fueling intergroup conflict and justifying exclusion and violence. These dynamics are alarmingly pervasive, spanning from the extreme historical examples of genocides against minorities in Germany and Rwanda to the ongoing violence and rhetoric targeting migrants in the US and Europe. While concepts like hate speech and fear speech have been explored in existing literature, they capture only part of this broader and more nuanced dynamic which can often be harder to detect, particularly in online speech and propaganda. To address this challenge, we introduce a novel computational framework that leverages large language models (LLMs) to quantify othering across diverse contexts, extending beyond traditional linguistic indicators of hostility. Applying the model to real-world data from Telegram war bloggers and political discussions on Gab reveals how othering escalates during conflicts, interacts with moral language, and garners significant attention, particularly during periods of crisis. Our framework, designed to offer deeper insights into othering dynamics, combines with a rapid adaptation process to provide essential tools for mitigating othering's adverse impacts on social cohesion.[1]

Introduction

In times of crisis, people turn to social media for information to help them make sense of events. Consequently, social media platforms can significantly influence individual perceptions and understanding of reality. This dynamic creates a strong incentive for various actors to manipulate public perceptions through online messaging and propaganda. Tactics that frame certain groups as separate and inherently dangerous are particularly effective at evoking strong emotional responses (Saha et al. 2021), often contributing to radicalization and intergroup conflict (Cervone, Augoustinos, and Maass 2021). This rhetoric spans from explicit hate speech and dehumanization (Buyse 2014; Kennedy et al. 2023) to subtler forms of fear speech (Buyse 2014; Saha et al. 2021; Schulze, Müller, and Lenz 2023; Saha et al. 2023).

While previous research studied some manifestations of this dynamic, like hate speech (Basile et al. 2019; Waseem and Hovy 2016; Kennedy et al. 2018) and fear speech (Saha et al. 2021, 2023), these represent just pieces of the broader social process of *othering*. Othering involves the systematic construction of an outgroup—the other—as fundamentally different and ultimately threatening to the ingroup. The goal of othering is to exclude and marginalize certain groups based on arbitrary characteristics like race, ethnicity or religion, ultimately establishing a self-sustaining group identity-based 'us-versus-them' mentality.

Historically, othering has been used to justify extreme measures and perpetuate cycles of violence and exclusion. Across various societies and periods, outgroups have been constructed as threats to the social, cultural, or political stability of the ingroup, legitimizing repression and violence. In Nazi Germany, discursive strategies mobilized hatred against Jews, Romani people, and homosexuals, reducing them to targets of mass extermination based solely on their group membership (Reicher, Haslam, and Rath 2008). Similarly, Stalinist terror was framed as a moral battle against an evil 'other', such as "wreckers, diversionists, and spies," with repression framed as necessary for the preservation of the state (Gerwarth 2007). In India, Hindutva organizations invoked narratives of Hindu tolerance to justify violence against Muslims, blaming conflicts on supposed Muslim intolerance (Kaur 2005). In recent years, social media has increasingly been used to spread narratives depicting immigrants and ethnic minorities as threats to national and cultural values (Nortio et al. 2021). In one particularly egregious example, Facebook was used to incite deadly riots and genocidal behavior in Myanmar against the Rohingya minority (Yue 2019). Ultimately, this process——driven by persistent themes of intergroup conflict and perceived threats—endures across communities and time, adapting to the evolving needs of the ingroup, serving as a self-sustaining force of exclusion. As such, tracking and quantifying this social dynamic on online platforms is crucial. While existing research has extensively explored explicit manifestations of intergroup conflict like hate speech and fear speech (Cervone, Augoustinos, and Maass 2021; Saha et al. 2021), these expressions are only part of the broader process of othering. Our work explores the subtler, lesser-understood mechanisms that sustain these expressions, shedding light on their nature and their interactions with moral language and attention mechanisms.

[1]Code to reproduce our models and findings available at https://anonymous.4open.science/r/othering_language-68FF/

We ground our analysis in sociological theory, defining 'othering' as the process of depicting a group as fundamentally different from one's own (Reicher, Haslam, and Rath 2008; Cikara 2015; Jetten, Spears, and Manstead 1997), this group self-talk is often characterized by the positive portrayal of one's own group (i.e., the ingroup) and the negative portrayal of the other group (i.e., the outgroup). This social dynamic marginalizes, excludes, or discriminates against the outgroup based on arbitrary or perceived differences, such as race, religion, or ethnicity, reinforcing social hierarchies and justifying unequal treatment (Reicher, Haslam, and Rath 2008; Duckitt 2003). We introduce a taxonomy of othering language and show that it subsumes and extends commonly used text indicators of intergroup conflict.

Additionally, we introduce an innovative computational framework that leverages large language models (LLMs) as classifiers and enables their rapid adaptation to new domains.After validating a model of othering, we use it to explore the language of intergroup conflict in real-world scenarios. We analyze a corpus of messages posted on Telegram by Russian and Ukrainian war bloggers during the ongoing war between Russia and Ukraine, as well as a corpus of messages posted on the social media platform Gab. We explore the following research questions:

1. How does the use of othering by Russian and Ukrainian war bloggers on Telegram change over the course of the war?
2. How do moral language and othering interact, especially in the expressions of intergroup conflict?
3. How does constructing and reinforcing the image of a target group as the 'other' affect social attention?
4. Does othering intensify during times of crisis, and in what ways are these behaviors more strongly rewarded?

Our analysis reveals the amplification of othering language in polarized online environments and its tendency to attract attention, especially during crises. While we find that othering language is often moralized across groups, its asymmetrical use by Russian war bloggers highlights its distinct utility in propaganda. By exploring these dimensions, we demonstrate how othering underpins more overt expressions like fear speech, hate speech, and exclusionary practices, offering crucial insights for developing strategies to counteract othering and mitigate its impact on social cohesion.

Related Work and Background

The Language of Intergroup Conflict Intergroup conflict often drives violence by framing outgroups as existential threats to the ingroup. From Nazi Germany to modern online spaces, fear-mongering and hateful rhetoric radicalize populations by portraying outgroups as dangerous and immoral, justifying hostility and violence (Greipl, Rothut, and Schulze 2022; Reicher, Haslam, and Rath 2008).

Such conflict escalates during crises—pandemics, financial collapses, or political upheaval—when fears are externalized and outgroups are scapegoated for societal problems. Economic hardship heightens competition and prejudice, as seen during the Great Depression. Similarly, the 2015 European Refugee Crisis saw the rise of exclusionary rhetoric, driven by political and social tensions (Pettersson and Sakki 2017). Misperceptions of outgroup hostility further exacerbate tensions, with groups mistakenly believing the other supports violence, as seen in the 2021 Israeli-Palestinian conflict. However, corrective interventions have shown promise in reducing these tensions (Nir et al. 2023).

A key psychological mechanism driving intergroup conflict is the perception of outgroup threat——the belief that outgroups endanger the ingroup's identity or very existence. This dynamic was particularly evident during the COVID-19 pandemic, where heightened awareness of mortality intensified xenophobia (Esses and Hamilton 2021).

Ultimately, these often-manufactured perceptions of threat can escalate intergroup conflict, leading to scapegoating that legitimizes violence and perpetuates hatred. This cycle frequently results in real-world violence, reinforcing instability and deepening social divisions (Fink 2018; Warofka 2018).

Computational scientists often operationalize the language of intergroup conflict through two key concepts: hate speech and fear speech. Hate speech refers to expressions intended to insult, degrade, or incite hostility toward a group based on attributes like race, religion, or gender (Mathew et al. 2021). It promotes explicit hostility typically through vilification and dehumanization (Basile et al. 2019; Waseem and Hovy 2016; Kennedy et al. 2018). In contrast, fear speech invokes existential fear, portraying the target group as a fundamental threat to the ingroup's survival, culture, or identity (Buyse 2014; Saha et al. 2021, 2023). Both forms of speech reinforce an 'us-versus-them' mentality, either by inciting hostility or by instilling fear, emphasizing the danger the outgroup poses to the ingroup's way of life. Together, they contribute to the broader process of othering, where groups are marginalized or excluded based on perceived differences (Reicher, Haslam, and Rath 2008; Cikara 2015; Jetten, Spears, and Manstead 1997)."

Social Mechanisms of Othering Othering language extends beyond specific expressions like hate speech and fear speech; it encompasses the broader sociological process of constructing an outgroup and excluding individuals based on race, religion, or ethnicity. This process has been observed throughout history, from the Holocaust to contemporary political discourse on immigration (Reicher, Haslam, and Rath 2008). Understanding these mechanisms is essential for mitigating their effects and preventing intergroup conflict.

We base our understanding of othering on Reicher's model (Reicher, Haslam, and Rath 2008), which conceptualizes hate as emerging from a continuous process of othering. As illustrated in Figure 1, extreme violence and genocide are driven by a distorted perception of group identity, where individuals are targeted solely for belonging to an outgroup (Reicher, Haslam, and Rath 2008; Kaur 2005). In this process, even disavowing one's group identity offers no protection, as othering reduces individuals to their perceived group membership, overriding personal actions or beliefs (Reicher, Haslam, and Rath 2008; Duckitt 2003).

The model outlines five key steps in the development of

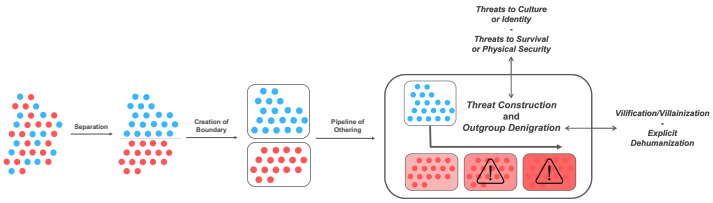


Figure 1: Conceptualization of the othering process. Othering starts with the separation of ingroup and outgroup members, the creation of symbolic boundaries, and the subsequent pipeline of othering. Through the construction and affirmation of perceived threats, the outgroup is increasingly framed as a threat.

collective hate: (i) creating a cohesive ingroup, (ii) excluding specific populations, (iii) framing the outgroup as a threat to the ingroup's existence, (iv) portraying the ingroup as virtuous, and (v) celebrating the outgroup's destruction as a defense of ingroup virtue (Reicher, Haslam, and Rath 2008). Central to this model is the formation of both ingroup and outgroup boundaries, which together play a critical role in fostering hostility and ultimately perceiving the outgroup as a threat (Reicher, Haslam, and Rath 2008). This dynamic is evident in ideologies like Nazism, where outgroups were framed as existential threats, justifying violence as moral and necessary for 'cleansing' and protecting the ingroup's values (Kennedy et al. 2023; Koonz 2003). Similarly, contemporary threats to the ingroup—whether economic or cultural—are often constructed to legitimize hostility. These identity narratives are not static but actively constructed, with the perceived danger of outgroups adapting to the fears of the ingroup (Reicher, Haslam, and Rath 2008).

The intersection of othering and moralized language is crucial for understanding how narratives of intergroup conflict are constructed. Moral language often serves as a powerful tool for legitimizing exclusion and violence (Fiske and Rai 2014; Kennedy et al. 2023). According to the *Moralized Threat Hypothesis*, extreme expressions of prejudice are frequently driven by the belief that the outgroup has committed a moral transgression, making violence not only justified but also morally righteous (Hoover et al. 2021). Consequently, we expect to find strong correlations between othering language—where a group is systematically framed as an enemy—and moral language.

Moreover, othering's role in capturing social attention demands further study. Narratives built around othering often attract disproportionate attention in online spaces, as the portrayal of an outgroup as a threat captures the audience's focus (Saha et al. 2023). This heightened attention amplifies the spread and impact of these narratives, particularly during times of crisis or polarization. Understanding this dynamic is essential for grasping how othering influences public discourse and fosters division

LLMs and In-Context Learning In-context learning enables LLMs to perform tasks by leveraging a few input-label pairs, or demonstrations, without requiring gradient updates, and has often outperformed zero-shot learning across numerous tasks (Zhao et al. 2021). However, its success is influenced by factors such as task complexity, the quality of provided examples, and the number of demonstrations. Our approach, which uses the *system prompt* to guide an LLM fine-tuned on one domain to adapt to another, offers a similar but distinct method for applying LLMs to unseen and untrained data.

Methods

Data

Russia-Ukraine war bloggers We use posts collected from Russian-oriented and Ukrainian-oriented Telegram channels (Theisen et al. 2022), spanning from October 2015 to August 2023. This dataset, which was gathered using both an expert-curated list and snowballing methods, includes 989 channels and over 9.67 million posts, primarily in Ukrainian and Russian. This data is released alongside this paper to support our findings and compel further research.
Labeling Telegram Channels: Telegram, a messaging app supporting private/public group interactions and one-way broadcasts via channels, has become a stronghold of a 'free' Internet in Russia. Evading bans affecting other major platforms like like Facebook and TikTok since 2020, Telegram has emerged as a key platform for military (war) bloggers and a primary source of news for both Ukrainians and Russians during the Russia-Ukraine war (Oleinik 2024). We construct a network of Telegram channels within the war bloggers corpus, where a directed link with weight w connects channel A to channel B if A references or forwards B's posts w times during the period. The network, Figure 11, shows the channels and connections between them. The global structure of information sharing shows roughly two clusters, as expected of groups in conflict.

We manually labeled 100 random channels as "pro-Russia', 'pro-Ukraine,' or 'Other' based on their bios and recent posts, then used these as seeds for a label propagation algorithm (Garza and Schaeffer 2019) to categorize the network. To validate, we reviewed 100 randomly selected channels from each group. The final dataset includes 243 pro-Ukrainian channels (4.2 million posts) and 325 pro-Russian channels (4.4 million posts) from October 2015 to August 2023, though messages prior to late 2021 are sparse.

Gab Corpus Introduced in prior work by Saha et al. (Saha et al. 2023), this corpus contains 9,441 text posts from the popular 'alt-tech' social media platform Gab (Dehghan and Nagappa 2022). Gab, which is popular with political conservatives in US, hosts discussions often revolving around issues of race, immigration, and national identity. Despite the absence of direct physical conflict, the rhetoric on Gab is steeped in othering narratives, making it a platform of choice for studying hate speech and fear speech (Kennedy et al. 2018; Saha et al. 2023). Each post was manually classified by humans annotators into one or more categories: 'normal,' 'fear speech,' or 'hate speech,' with some posts receiving multiple labels. We detail the definitions used for fear speech and hate speech in a later section on othering language's relationship to expressions of intergroup conflict. Overall, 44.8% of the posts are labeled as normal, 19.7% as fear speech, and 42.4% as hate speech

A Model of Othering

We develop a flexible, LLM-based model to recognize othering language in text and show how it can be rapidly adapted to new domains.

Taxonomy of Othering Othering is a group self-talk process that helps shape group's conceptualization of itself as good and virtuous and the other group as inherently evil and dangerous. When talking about itself, i.e., the ingroup, the group uses fear-laden speech (Lerman et al. 2024), which serves to bind the group together, often in response to a perceived threat from the outgroup. When talking about the other, i.e., the outgroup, othering manifests itself through animosity and hostility (Stephan, Ybarra, and Rios 2015; Joffe 1999). To capture the various dimensions of othering, we define and provide translated examples from Russian war bloggers for four categories of language linked to the othering process: the first two categories address perceived threats to the ingroup, while the latter two focus on the demonization of outgroups.

Threats to Culture or Identity arise when the outgroup is framed as a danger to the ingroup's cultural or social survival, challenging its values, language or traditions (Wohl, Branscombe, and Reysen 2010; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015; Joffe 1999): *"The erosion of the Russian language in Ukrainian schools: Ukrainian policymakers pushing to erase the Russian tongue risk severing the threads that weave together our history."*

Threats to Survival or Physical Security involve portraying the outgroup as an existential menace to the ingroup's physical well-being, thereby justifying preemptive hostility (Wohl, Branscombe, and Reysen 2010; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015; Joffe 1999): *"Zelensky's regime has accumulated 30 tons of plutonium and 40 tons of enriched uranium at the Zaporizhia NPP [...] the regime really is on the verge of creating its own nuclear bomb! And hundreds of 'dirty' bombs can be made from such a quantity of radioactive material!"*

Vilification/Villainization casts the outgroup as inherently evil or immoral, which in turn validates resistance and aggression (Joffe 1999; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015): *"Because these Ukronazi girls can fight only by hiding behind hostages. All their courage went down the drain in chants and slogans like 'hang the Muscovite.' But when the Russians came, they shit themselves, just like their Bandera."*

Explicit Dehumanization represents the most extreme form of othering, where the outgroup is compared to animals, objects or spirits, paving the way for extreme violence (Joffe 1999; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015): *"These are zombies, who may have been brothers before, but over the past 8 years, from the bite of Nazism and Banderization, they have turned into non-humans. That is why our army calls on all brothers to lay down their arms, so that we can distinguish a brother from an infected zombie, who can only bite and infect."*

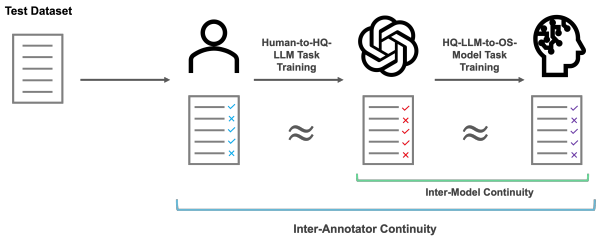


Figure 2: Artificial Annotator Alignment process. Human-annotated data is first used to train a high-quality LLM. The HQ-LLM's annotations are compared with human annotations for alignment, and then the HQ-LLM is used to annotate a larger dataset. Finally, an open-source model is trained using the HQ-LLM-annotated data, optimizing both effectiveness and cost-efficiency.

These classes are integral to understanding how outgroups are systematically constructed, legitimizing their exclusion and violence.

Artificial Annotator Alignment Process We train a classifier to recognize othering language, using an "Artificial Annotator Alignment" process inspired by knowledge distillation (Gou et al. 2021) to efficiently train the model and adapt it to new domains. Our alignment process, shown in Figure 2, essentially "trains" LLMs as annotators by requiring them to be consistent with human annotators. Fist, human annotators label a small subset of the data. This subset is then annotated by a "high-quality" (HQ) LLM like ChatGPT-4. If the HQ-LLM's annotations closely align with those of the human annotators, we proceed to have the HQ-LLM annotate more data to effectively train an open-source LLM (OS-LLM), e.g., Llama3, as an annotator. This strategy is driven by the goal of maximizing effectiveness and cost: utilizing a combination of an HQ-LLM and OS-LLM to avoid the prohibitive costs of large-scale annotation.

This approach is guided by two core principles: (1) The initial, high-quality (HQ) LLM must produce annotations that closely resemble those made by humans; (2) The open-source (OS) LLM trained on the HQ-LLM-annotated data must achieve performance on par with the HQ-LLM when evaluated against the human-annotated data (which is held-out throughout the process for evaluation). To adhere to the first principle, we assess the continuity between human annotations and the HQ-LLM using both standard machine learning metrics (accuracy, F1-score, etc.) and inter-annotator agreement metrics. This approach allows us to evaluate the HQ-LLM both as a classifier and as an artificial annotator, ensuring its consistency with human annotations. If both sets of evaluations yield optimal results, we consider the HQ-LLM a reliable proxy for human annotation. We then proceed to use the HQ-LLM to annotate a substantially larger portion of the dataset.

To adhere to the second principle, after training the target OS-LLM on the data annotated by the HQ-LLM, we test its performance on the human-annotated dataset, once again using both sets of metrics to ensure that its classification metrics do not degrade from the HQ-LLM (ensuring inter-model

continuity) and that its classifications are consistent with human annotations (ensuring inter-annotator continuity). This two-step validation ensures that the target model OS-LLM not only replicates the quality of the HQ-LLM's annotations but also aligns with human judgment, thereby confirming its effectiveness in the domain.

Data Annotation Our analysis began with two datasets: messages from Russian war bloggers and from Ukrainian war bloggers. For each dataset, we used a combination of random post selection and keyword-based upsampling, targeting phrases and coded terms of denigration specific to each context (e.g., "Ukronazis" in the Russian data).

Human Annotation: The approach yielded 316 posts for the Russian data, which were labeled by six human annotators with overlapping annotations. A similar process was applied to the Ukrainian dataset, but with two annotators, based on the strong performance observed in the Russian workflow. Inter-annotator agreement, shown in Tables 6 and 9 for the Russian and Ukrainian data, is consistent with benchmarks in similar studies (Saha et al. 2021, 2023). Final classification counts for each domain were determined through majority vote, as summarized in Tables 5 and 9.

High-Quality LLM Annotation: Following independent human annotations, we used ChatGPT-4o (HQ-LLM) to annotate the same examples using prompts tailored to the specific context (prompts available in our GitHub repository), outputting a dictionary for each post: "'Threats to Culture or Identity': 1, 'Threats to Survival or Physical Security': 0, 'Vilification/Villainization': 1, 'Explicit Dehumanization': 0, 'None': 0", along with an explanation. For example, "The text describes local Nazis desecrating a historic Russian cemetery, in a way that represents a threat to cultural identity and vilifies the opposing group." These explanations were crucial for understanding the rationale behind each annotation and for testing our Rapid Domain Adaptation method later. The annotations were validated using metrics, such as Cohen's Kappa, treating ChatGPT-4o as an additional annotator. The results, detailed in tables 10, show that the ChatGPT-4o annotations were reliable and consistent across domains. In total, ChatGPT-4o annotated 20,000 posts (10,000 from each dataset) at a cost of approximately $70 USD, significantly lower than human annotation costs while remaining consistent with human annotators.

Training Models Next, we trained three different models (OS-LLMs) on the ChatGPT-4o-annotated (HQ-LLM) data: Mistral, LLaMA3-8b-Instruct, and LLaMA2. Each model was trained separately on three datasets: Russian-only, Ukrainian-only, and a combined Russian-and-Ukrainian dataset. To ensure a fair evaluation across different domains, we withheld common test and validation sets for each dataset, preventing data leakage that could give a model an unfair advantage. Each model was trained for 5 epochs on an NVIDIA Quadro RTX 8000 GPU using a learning rate of 1e-5 and followed a 0.7:0.1:0.2 split for training, validation, and testing. The best-performing model from each run was selected based on the best F1 score on the validation set.

After training, we evaluated each model across three domains: Ukrainian-only, Russian-only, and combined Russian-and-Ukrainian. The exact performance metrics for each domain-model pair are available in our repository, with the best-performing model—LLaMA3-8b-Instruct—detailed in Table 1. Our evaluation shows that the model adheres to principle 2, exhibiting minimal degradation in ML-based metrics compared to ChatGPT-4o and remaining consistent with human annotators when assessed as an artificial annotator.

Models trained on one domain perform best within that domain but struggle to generalize to other domains. In contrast, models trained on multiple domains demonstrate better cross-domain generalization, though they don't match the performance of single-domain models within their specific domain.

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.84	0.89	0.86
Threats to Survival or Physical Security	0.74	0.84	0.80
Vilification/Villainization	0.81	0.91	0.89
Explicit Dehumanization	0.84	0.89	0.88
None	0.84	0.90	0.87

Table 1: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between Majority Vote and LLM on Russian Data.

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.81	0.88	0.89
Threats to Survival or Physical Security	0.72	0.87	0.82
Vilification/Villainization	0.80	0.87	0.86
Explicit Dehumanization	0.82	0.89	0.88
None	0.83	0.88	0.84

Table 2: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between Majority Vote and LLM on Ukrainian Data.

Rapid Domain Adaptation Adapting models to new domains presents a significant challenge in classification tasks. To study othering across multiple domains in a cost-effective manner, we test whether our models could effectively transfer knowledge from one domain to another. LLMs are well-suited for this task due to their (1) inherent complexity and (2) the pseudo-world mapping generated through extensive pretraining on vast corpora. To leverage this power, we employ two techniques that adapt the classifier to new, unseen data: system prompt steering and logit disambiguation. We name this approach Rapid Domain Adaptation (RDA).

System Prompt Steering: The first component of our RDA system involves manipulating the system prompt for the trained model. A system prompt provides the model with initial instructions, guiding how it processes and classifies incoming data. While in-context learning offers some benefits, simply appending context to new data only minimally improves performance. However, a well-crafted system prompt, designed to steer the model's reasoning, led to significant improvements in new domains by explicitly framing the new domain (this logic is illustrated in Figure 12) in relation to the model's prior training, we achieved notable enhancements in domain adaptation.

Logit Disambiguation: The second component of our RDA system involves exposing the logits for each class, rather than simply assigning a binary label of 0 or 1. Logits represent the model's confidence in its predictions, indicating the likelihood of a particular token being chosen. Since our task is to classify messages using either 1 or 0 for each class, we can expose the logits for this classification token and then using confidence thresholds for each class, fine-tuning them specifically for new domains. This approach is especially useful when the new domain features significantly different language or topics compared to the original training domain. As shown in Figure 10, by disambiguating the logits and adjusting confidence thresholds, we can better adapt the model to the new domain.

Overall, these components — system prompt steering and logit disambiguation — work together to enable rapid and reliable domain adaptation, leveraging modern LLMs to effectively handle drastic shifts in domain context.

RDA Evaluation: We evaluated our RDA system across all cross-domain combinations (e.g., a model trained on Russian war blogger data performing on Ukrainian war blogger data), with detailed results available in our repository. We first compared system prompt steering to two baseline approaches: no added context and traditional in-context learning, where context is simply appended to the new prompting data. Table 13 shows the F1 scores for different domain transitions. Figure 3 visualizes the substantial performance gains across metrics when a model trained on Russian data applied to Gab, which contains messages posted on a different platform, in a different language (English), and in a different cultural and geopolitical context (representing the most substantial domain transition). These results demonstrate the effectiveness of our RDA system, especially in models that had no prior exposure to the new domain.

Prompting Type	Accuracy	F1 Score	Precision	Recall
No Additional Prompt	0.55	0.53	0.83	0.56
In-Context Learning	0.61	0.63	0.84	0.65
System Prompt	0.69	0.76	**0.90**	0.70
RDA	**0.71**	**0.77**	**0.90**	**0.71**

Table 3: Performance Comparison of Different Prompting Types: Accuracy, F1 Score, Precision, and Recall.

Relationship to Language of Intergroup Conflict

We examine the relationship between othering language and other widely studied expressions of intergroup conflict, such as fear speech and hate speech, using the annotated Gab corpus. As illustrated in Fig. 4, fear speech and hate speech reflect expressions of othering, but they do not fully encompass it. Instead, they function as partial components of the broader process. This underscores the asymmetry suggested by our model, which posits that othering language subsumes, but is not limited to, specific expressions like fear speech and hate speech. Additionally, we consider the practical implications for content moderation by evaluating how well current toxicity classifiers detect othering language alongside fear speech and hate speech. This analysis highlights the limitations of existing classification tools in addressing the full spectrum of othering language.

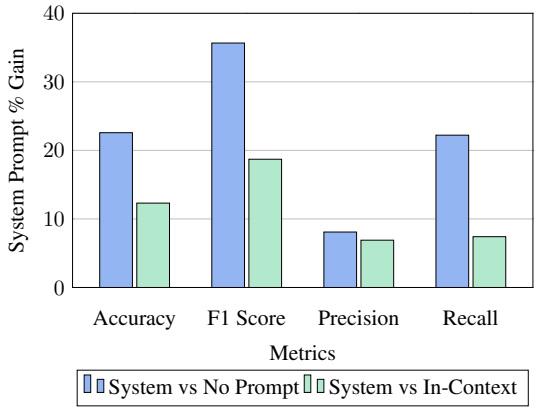


Figure 3: System Prompt % Gain Compared to No Additional Prompt and In-Context Learning Across Metrics (Accuracy, F1 Score, Precision, Recall).

Fear Speech Fear speech, defined as expressions that instill existential fear of a target group (often based on attributes such as race, religion, or gender) (Buyse 2014), is significantly associated with othering language. Specifically, the probability that fear speech contains 'Vilification/Villainization' is 68.9%, and the probability of containing 'Threats to Culture or Identity' is 50.5%. It has a weaker association with 'Threats to Survival or Physical Security' at 20.3%. These connections reflect the nature of fear speech in both evoking existential fear and subtly vilifying the outgroup, such as in messages like "They will destroy our way of life unless we stop them.' Moreover, fear speech shows an asymmetric relationship to othering: 88.9% of fear speech instances involve othering, but only 24.2% of othering messages are classified as fear speech.

Hate Speech Hate speech is language used to express hatred toward a targeted individual or group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on attributes such as race, religion, or gender (Mathew et al. 2021), and is typically the most explicit form of othering. Our analysis shows strong associations with 'Vilification/Villainization' (74.1%), 'Explicit Dehumanization' (37.3%), and 'Threats to Culture or Identity' (32.3%). These connections underscore hate speech's dual role both denigrating the outgroup and framing it as a threat.

Hate speech also demonstrates an asymmetric relationship with othering language: approximately 87.4% of hate speech involves othering, but only 51.1% of messages containing othering language are classified as hate speech.

Toxicity Finally, we analyze the relationship to toxicity. Using the Detoxify classifier[2], which rates text on a scale from 0 to 1 (with scores above 0.5 considered toxic), we find the following average toxicity scores: fear speech averages 0.46, while hate speech scores higher at 0.65. For othering content, excluding Explicit Dehumanization, the av-

[2]https://github.com/unitaryai/detoxify

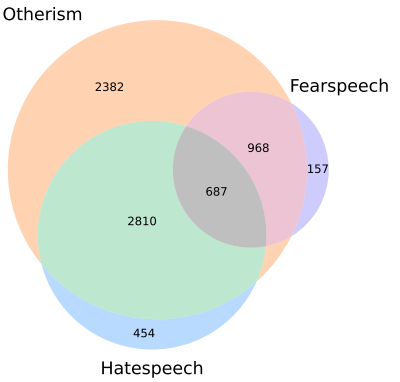


Figure 4: Venn diagram showing overlap between othering, fear speech, and hate speech in the Gab corpus. The diagram reveals that while fear speech and hate speech often co-occur with othering, many instances of othering occur without these explicit forms of conflict language.

erage toxicity is 0.42 and increases to 0.53 when Explicit Dehumanization is included (which has a notably high average toxicity of 0.81). These findings highlight the broad spectrum of toxicity within othering rhetoric and emphasize the need for more effective detection tools.

Results

We label the corpus of messages posted on Telegram by Russian and Ukrainian war bloggers for othering language, focusing on the period from late 2021 to August of 2023. This period of war was characterized by high conflict and animosity between the groups. We also label a corpus of messages posted on Gab, which is often favored by far-right users within the US.

Othering during the Russia-Ukraine War

Figure 5 shows that othering rhetoric fluctuated throughout the conflict, rising after key events. We define key events as significant events during the war that also became prominent discussion topics within their respective communities (i.e., Russian and Ukrainian war bloggers separately). Here, we utilize the events enumerated in Tables 15 and 16, which were compiled using domain knowledge in conjunction with methods from previous work (Gerard et al. 2024). Russian war bloggers used more othering language overall than Ukrainian war bloggers, with the exception of dehumanization, which was lower overall. We observe that othering tends to respond to current events, political rhetoric, on-the-ground developments, and international incidents viewed as important to that specific community. For instance, following the invasion on February 24, 2022, there was a noticeable rise in othering language among both Ukrainian and Russian war bloggers. However, after the EU sanctions on Russian oil, an uptick in othering language occurred exclusively among Russian war bloggers, primarily in the form of *Threats to Culture or Identity* as they framed themselves as victims. In future work, we aim to automate event detection using change point analysis and quantify the causal

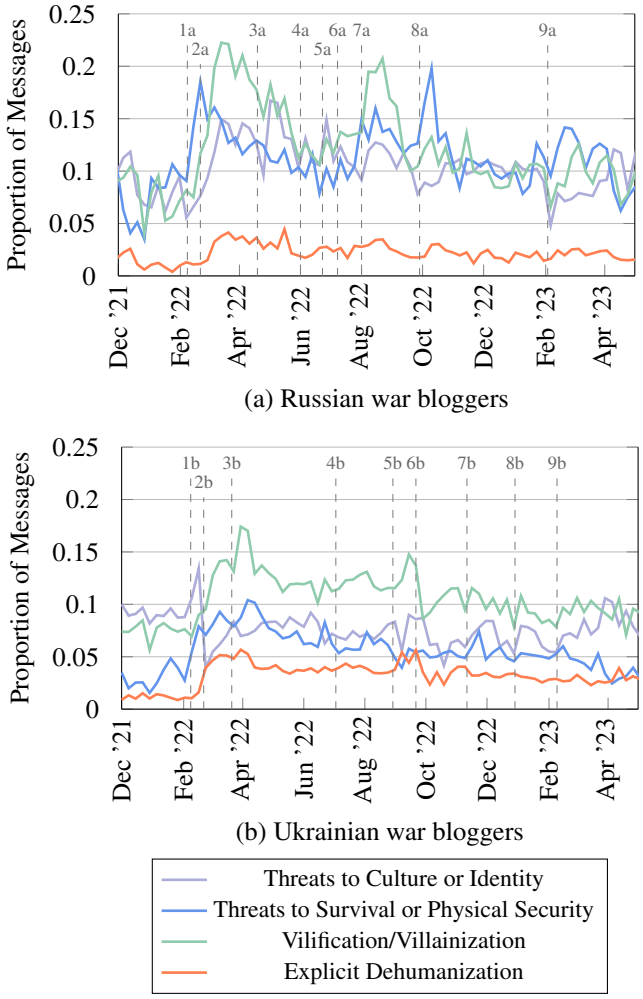


Figure 5: Temporal trends in the proportion of messages with othering language posted by (a) **Russian war bloggers** and (b) **Ukrainian war bloggers** from December 2021 to May 2023. The four classes of othering language are: Threats to Culture or Identity, Threats to Survival or Physical Security, Vilification/Villainization, and Explicit Dehumanization.

relationship between these events and the prevalence of othering language.

The Moral Language of Othering

We explore the interaction between othering and moral language on Telegram and Gab. For both platforms, we label moral values expressed in text using a model trained to recognize moral language (Trager et al. 2022). The model identifies the moral foundations of human intuitive ethics, such as valuing of purity, respect for authority, equality (fairness), group loyalty, and care or protection of the more vulnerable (Graham et al. 2013).

Moralized Othering on Telegram We begin by analyzing messages posted by war bloggers on Telegram, using a chi-squared test to explore the relationship between moral language in messages containing othering language

versus those without it. The chi-squared test helps determine whether the association between moral language and othering is statistically significant, rather than occurring by chance. The results indicate a strong, statistically significant connection ($p < 0.001$), demonstrating that moral language is more likely to co-occur with othering language than with non-othering language. This suggests that moral framing is frequently used to justify or intensify othering language, supporting the Moralized Threat Hypothesis (Hoover et al. 2021).

Next, we analyzed the use of moral language by Russian and Ukrainian war bloggers across different othering categories to identify differences in their moral framing strategies. Figure 7 highlights the significant variations in the moral values expressed by each side when using othering language, based on log-odds ratios. We also examined the interaction between specific moral language categories and individual othering classes. As shown in Figure 14, the two groups also differ significantly in their use of moral language within othering rhetoric, with these differences confirmed by two-proportion z-tests ($p < 0.001$).

Explicit forms of othering, such as *Explicit Dehumanization*, are the least associated with moral language in both groups, likely due to their overtly aggressive nature, which doesn't align well with broader moral frameworks (Saha et al. 2023). However, for Russian war bloggers, the strongest association between morality and othering is found in the purity moral frame within *Explicit Dehumanization*. This suggests that while Russians use moral language less frequently with dehumanization, when they do, it is often tied to purity, reflecting popular narratives portraying Ukrainians (and the West) as 'puppets' or 'zombies' corrupting Russian values. Meanwhile, for Ukrainian war bloggers, the most morally charged category is *Threats to Survival or Physical Safety*, most closely associated with care, which aligns with the context of Russia's invasion.

Overall, both groups display similar trends, but the use of moral language reveals strategic differences. Russian bloggers emphasize purity and cultural threats, reinforcing existential threat and victimization (Geissler et al. 2023), while Ukrainian bloggers focus on care in response to physical threats. Our analysis shows that moral language and othering are deeply intertwined, with Russian bloggers heavily relying on moralized language to justify intergroup prejudice. This suggests moralized othering is an effective propaganda tool, though further research is needed to understand its role in polarized environments.

Moralized Othering on Gab Gab provides an additional lens for examining the interplay between moral language and othering in a polarized environment with lower levels of immediate conflict. Although not directly involved in intergroup violence, the platform's rhetoric is steeped in othering narratives within an American context. This makes Gab an ideal setting to further test our hypothesis in a context similarly shaped by divisive and exclusionary discourse, but without direct conflict.

As in the previous case study, we first applied the chi-squared test to confirm a significant difference in the use of moral language between messages with and without othering language ($p < 0.001$). We then evaluated the log-odds ratios of moral language in othering messages and the interaction between specific moral categories and othering types, as shown in Figures8 and 15. The most morally loaded othering category is *Threats to Survival and Physical Security*, primarily associated with care, similar to Ukrainians. *Threats to Culture or Identity* is linked to equality and loyalty, similar to Russians. *Explicit Dehumanization* is tied to purity but not to other moral values, unlike in Russians, where it had (weak) links to all moral values.

These results not only support the hypothesis that moral language and othering are closely intertwined but also reaffirm the findings from the previous case study, highlighting how moral language shapes and sustains othering narratives. They also offer a new lens on how different moral devices

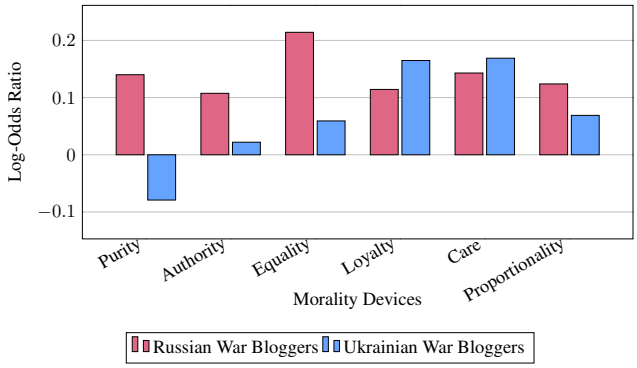


Figure 6: Log-odds ratios of morality devices for Russian and Ukrainian war bloggers, comparing the presence of moral language in messages with othering language versus those without. Ukrainian war bloggers are represented in light blue, and Russian war bloggers in light red.

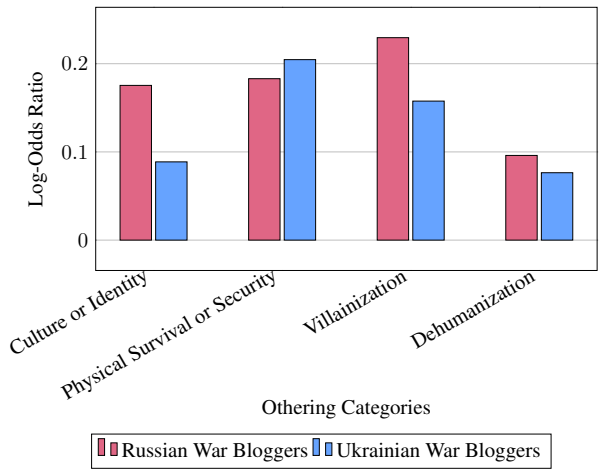


Figure 7: Log-odds ratios of morality use for Russian and Ukrainian war bloggers, comparing its use in messages with specific othering language use versus those without. Ukrainian war bloggers are represented in light blue, and Russian war bloggers are represented in light red.

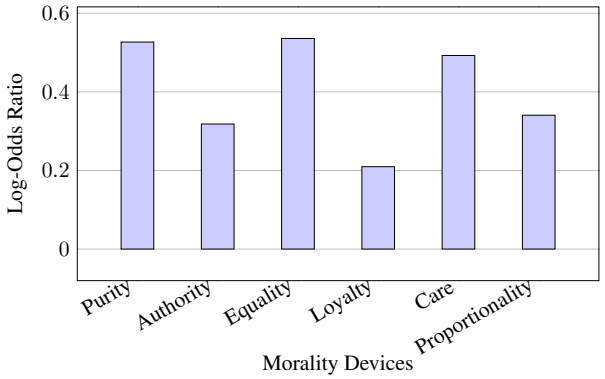


Figure 8: Log-odds ratios of morality devices for Gab, comparing the presence of moral language in messages with othering language versus those without.

are strategically used to appeal to different audiences, underscoring the need for further research into how these narratives are constructed and spread in polarized environments.

Othering and Attention

Next, we explore the relationship between online attention and the use of othering language in messages by Russian and Ukrainian war bloggers.

Network Centrality We measure channel centrality in the reference network of war blogger channels (see Methods) using degree centrality and eigenvector centrality. (For simplicity, we use an undirected version of this network.) Degree centrality reflects the influence distribution and communicative ability of nodes in the network and eigenvector centrality captures the positional importance of network nodes. We then calculate the Spearman correlation between channel centrality and its use of othering language (as a proportion of its messages).

The results, shown in Table 4, reveal statistically significant correlations for both degree centrality and eigenvector centrality and the use of othering language by both groups of war bloggers, with a stronger correlation among Russian war bloggers. This suggests that war bloggers who use othering language occupy more influential positions within the Telegram network, though the reasons remain unclear. It could be that existing opinion leaders are more inclined to use inflammatory othering language, or that such language attracts more attention, rewarding those who use it. This finding does not disentangle these possibilities.

Message Views To further investigate the link between othering and attention, we focus on the messages themselves. We normalize each message's number of views by the channel's typical viewership using Z-Score normalization. As shown in Figure 9, messages containing othering language consistently garner more views than those without, a relationship confirmed by the Mann-Whitney U Test ($p < 0.001$), which demonstrates a statistically significant difference between the number of views of messages with and without othering language. These results show that othering messages attract more attention than regular messages.

Times of Crisis We define times of crisis as the time period immediately following significant events (one week) during the war that are popular discussion topics amongst the community (in this case Russian and Ukrainian war bloggers separately). Here, we utilize the events enumerated in Tables 15 and 16. Overall, we find that othering receives more attention immediately following a crisis.

As shown in Figure 14, the correlation between degree centrality and the proportion of messages containing othering increases markedly following events. For Russian war bloggers, there is also a significant rise in the correlation between eigenvector centrality and othering, while for Ukrainian war bloggers, there is a slight decrease. Additionally, Figure 13 illustrates a shift in viewership dynamics. For Ukrainian war bloggers, the gap in viewership between messages with and without othering widens dramatically during periods of heightened conflict, a statistically significant change confirmed by the Mann-Whitney U Test. In contrast, the viewership differential for Russian war bloggers remains relatively stable, even during crises. These findings suggest that while othering generally increases attention, its impact becomes particularly pronounced in times of crisis.

This analysis highlights that othering not only correlates with increased attention in online discourse but may also be structurally rewarded, especially during crises. Both network centrality and viewership metrics suggest that users engaging in othering are more likely to occupy influential positions within their networks and enjoy greater visibility. The sharp increase in attention during periods of heightened conflict, particularly among Ukrainian war bloggers, emphasizes the role of othering in driving engagement. These findings strongly suggest that othering functions as a powerful tool for capturing attention and influence in polarized environments, with its effects significantly amplified in times of crisis.

	Centrality Metric	
Community	Degree	Eigenvector
Russian	0.254	0.333
Ukrainian	0.128	0.147

Table 4: Centrality and othering messages. Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities. All correlations are significant at the $p < 0.01$ level.

Discussion and Conclusion

Our work presents a comprehensive taxonomy of othering language, grounded in sociological theory and contextualized alongside established concepts like hate speech and fear speech. This framework offers a structured approach to identifying and understanding othering language, clarifying its role in online discourse. We also introduce methods for efficiently training LLMs to detect othering language and transfer this knowledge to new domains, aiming to help advance future efforts in understanding and moderating it.

Our analysis highlights the dynamic nature of othering language and its responsiveness to real-world events. We establish a clear link between othering and moral language, reinforcing the Moralized Threat Hypothesis and showing

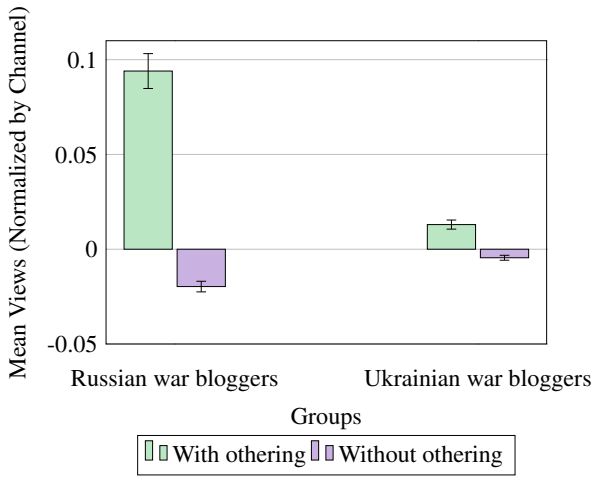


Figure 9: Comparison of mean views with and without othering (z-score channel-normalized). The bars represent the mean views for Russian and Ukrainian war bloggers, with and without othering, and the error bars indicate the standard error.

how moral framing can amplify harmful narratives. Additionally, we find a strong connection between othering language and social attention, with such messages receiving higher engagement, especially during times of crisis when othering intensifies and attracts disproportionately more attention.

Ethical Considerations

Our study used publicly accessible data from influential figures discussing the Russia-Ukraine war on Telegram, following the FAIR data principles. While we aim to advance computational analysis, we recognize the risks of misuse, such as evading classifiers or generating harmful language; applying our classifiers poses its own challenges: over-detection may suppress legitimate speech, while under-detection can spread harmful rhetoric. To ensure transparency and responsible use, we have made both the data and code publicly available, stressing the ethical application of these resources.

Limitations

Our analysis is centered on Russian and Ukrainian war bloggers operating in a highly charged war environment, which may impact the neutrality of the data. Additionally, while the study spans the first year and a half of the conflict, the war is ongoing, and future developments may alter the narrative dynamics. The centrality metrics we used are based on networks generated over the entire timeframe, which limits their dynamism compared to platforms like Twitter, where more edges are created. Finally, while our model allows for editing the system prompt in the RDA process, many other open-source models, such as Mistral, do not offer this flexibility, which may limit adaptability in similar studies.

In addition, biases could inadvertently find their way into our results. One source of bias is moral annotation process, which was trained on out-of-domain data. However, when doing comparative analysis, biases should largely cancel.

Summary and Future Work

This study analyzes how othering language is used by Russian and Ukrainian war bloggers on Telegram, highlighting its dynamic evolution in response to external events and its interaction with moral language and social attention. We develop a taxonomy for othering language and introduce methods for training LLMs to detect and transfer this knowledge to new domains. Our findings show that othering intensifies during crises, attracting more social attention and rewards, complicating the online discourse.

Future work should extend these methods to other conflicts and domains to test their generalizability, providing insights into how othering and social attention interact in polarized environments. Additionally, future research should explore how proximity to the outgroup shapes othering language. Specifically, core members may rely more on ideological and coded language to reinforce identity, while boundary members tend to adopt defensive and confrontational rhetoric, justifying aggression and emphasizing immediate threats. Understanding these dynamics could reveal further nuances in the strategic use of moralized othering.

References

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 54–63.

Buyse, A. 2014. Words of violence:" Fear speech," or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, 36(4): 779–797.

Cervone, C.; Augoustinos, M.; and Maass, A. 2021. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of language and social psychology*, 40(1): 80–101.

Cikara, M. 2015. Intergroup schadenfreude: Motivating participation in collective violence. *Current opinion in behavioral sciences*, 3: 12–17.

Dehghan, E.; and Nagappa, A. 2022. Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab Social. *Social Media+ Society*, 8(3): 20563051221113075.

Duckitt, J. 2003. Prejudice and intergroup hostility.

Esses, V. M.; and Hamilton, L. K. 2021. Xenophobia and anti-immigrant attitudes in the time of COVID-19. *Group Processes & Intergroup Relations*, 24(2): 253–259.

Fink, C. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs*, 71(1.5): 43–52.

Fiske, A. P.; and Rai, T. S. 2014. *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Garza, S. E.; and Schaeffer, S. E. 2019. Community detection with the label propagation algorithm: a survey. *Physica A: Statistical Mechanics and its Applications*, 534.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1): 35.

Gerard, P.; Volkova, S.; Penafiel, L.; Lerman, K.; and Weninger, T. 2024. Modeling Information Narrative Detection and Evolution on Telegram during the Russia-Ukraine War. arXiv:2409.07684.

Gerwarth, R. 2007. The Dictators: Hitler's Germany and Stalin's Russia.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Greipl, S.; Rothut, J.; and Schulze, M. 2022. Online Radicalization: The Role of Fear. *Journal of Online Behavior*, 10(3): 203–220.

Hoover, J.; Atari, M.; Mostafazadeh Davani, A.; Kennedy, B.; Portillo-Wightman, G.; Yeh, L.; and Dehghani, M. 2021. Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, 12(1): 4585.

Jetten, J.; Spears, R.; and Manstead, A. S. 1997. Strength of identification and intergroup differentiation: The influence of group norms. *European journal of social psychology*, 27(5): 603–609.

Joffe, H. 1999. *Risk and'the Other'*. Cambridge University Press.

Kaur, R. 2005. *Performative politics and the cultures of Hinduism: Public uses of religion in western India*. Anthem Press.

Kennedy, B.; Atari, M.; Davani, A. M.; Yeh, L.; Omrani, A.; Kim, Y.; Coombs, K.; Havaldar, S.; Portillo-Wightman, G.; Gonzalez, E.; et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Kennedy, B.; Golazizian, P.; Trager, J.; Atari, M.; Hoover, J.; Mostafazadeh Davani, A.; and Dehghani, M. 2023. The (moral) language of hate. *PNAS nexus*, 2(7): pgad210.

Koonz, C. 2003. *The nazi conscience*. Harvard University Press.

Lerman, K.; Feldman, D.; He, Z.; and Rao, A. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1): 8.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Nir, N.; Nassir, Y.; Hasson, Y.; and Halperin, E. 2023. Kill or be killed: Can correcting misperceptions of out-group hostility de-escalate a violent inter-group out-break? *European Journal of Social Psychology*, 53(5): 1004–1018.

Nortio, E.; Niska, M.; Renvik, T. A.; and Jasinskaja-Lahti, I. 2021. 'The nightmare of multiculturalism': Interpreting and deploying anti-immigration rhetoric in social media. *New Media & Society*, 23(3): 438–456.

Oleinik, A. 2024. Telegram channels covering Russia's invasion of Ukraine: a comparative analysis of large multilingual corpora. *Journal of Computational Social Science*, 1–24.

Pettersson, K.; and Sakki, I. 2017. Pray for the fatherland! Discursive and digital strategies at play in nationalist political blogging. *Qualitative Research in Psychology*, 14(3): 315–349.

Reicher, S.; Haslam, S. A.; and Rath, R. 2008. Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, 2(3): 1313–1344.

Saha, P.; Garimella, K.; Kalyan, N. K.; Pandey, S. K.; Meher, P. M.; Mathew, B.; and Mukherjee, A. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11): e2212270120.

Saha, P.; Mathew, B.; Garimella, K.; and Mukherjee, A. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *TheWebConf 2021*, 1110–1121.

Schulze, M.; Müller, H.; and Lenz, S. 2023. Fear-Based Messaging in Extremist Communication. *Journal of Communication Studies*, 12(1): 99–115.

Stephan, W. G.; Ybarra, O.; and Rios, K. 2015. Intergroup threat theory. In *Handbook of prejudice, stereotyping, and discrimination*, 255–278. Psychology Press.

Theisen, W.; Cedre, D. G.; Carmichael, Z.; Moreira, D.; Weninger, T.; and Scheirer, W. 2022. Motif Mining: Finding and Summarizing Remixed Image Content. arXiv:2203.08327.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazazian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M. 2022. The Moral Foundations Reddit Corpus. arXiv:2208.05545.

Warofka, A. 2018. An independent assessment of the human rights impact of Facebook in Myanmar. *Facebook Newsroom, November*, 5.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wohl, M. J.; Branscombe, N. R.; and Reysen, S. 2010. Perceiving your group's future to be in jeopardy: Extinction threat induces collective angst and the desire to strengthen the ingroup. *Personality and Social Psychology Bulletin*, 36(7): 898–910.

Yue, N. 2019. The "Weaponization" of Facebook in Myanmar: A Case for Corporate Criminal Liability. *Hastings LJ*, 71: 813.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, 12697–12706. PMLR.

Appendix

Category	Instance Counts
Threats to Culture or Identity	122
Threats to Survival or Physical Security	62
Vilification/Villainization	164
Explicit Dehumanization	77
None	78
Total Data Points	**316**

Table 5: Summary of human annotations for Russian war bloggers data.

Category	Krippendorff's	Fleiss'
Threats to Culture or Identity	0.72	0.72
Threats to Survival or Physical Security	0.62	0.62
Vilification/Villainization	0.70	0.73
Explicit Dehumanization	0.68	0.65
None	0.70	0.73

Table 6: Inter-Annotator Agreement: Krippendorff's Alpha and Fleiss' Kappa for Russian war bloggers data.

Category	Cohen's	Accuracy	F1
Threats to Culture or Identity	0.83	0.92	0.92
Threats to Survival/Security	0.75	0.82	0.80
Vilification/Villainization	0.80	0.90	0.90
Explicit Dehumanization	0.85	0.97	0.97
None	0.80	0.92	0.92

Table 7: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between majority vote and HQ-LLM (ChatGPT-4o) on Russian war bloggers data.

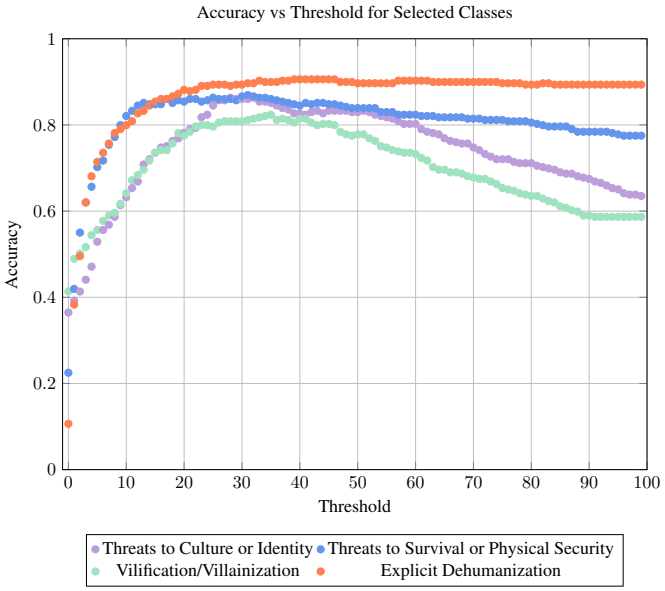


Figure 10: Scatter plot of accuracy vs. threshold for selected classes. This plot illustrates how adjusting confidence thresholds for classification logits affects accuracy across different classes in a new domain (Gab corpus), using a model initially trained on Russian war bloggers.

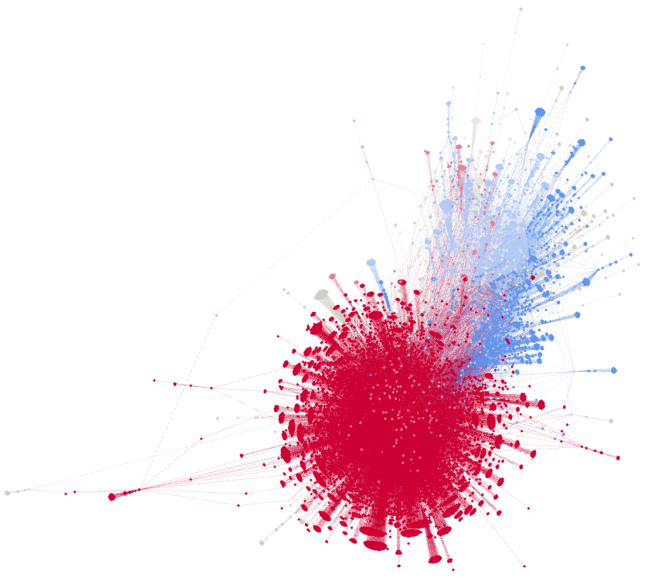


Figure 11: Visualization of a co-reference network based on message content and channel bios. Nodes are colored according to the stance indicated by their activity: Pro-Russian nodes are in red, Pro-Ukrainian nodes are in blue, and nodes not strongly affiliated to either nationality are in grey. This network was constructed by analyzing messages and bio information from various Telegram channels.

Category	Instance Counts
Threats to Culture or Identity	45
Threats to Survival or Physical Security	41
Vilification/Villainization	52
Explicit Dehumanization	32
None	87
Total Data Points	**212**

Table 8: Human-annotated gold set summary for Ukrainian war bloggers data.

Category	Cohen's
Threats to Culture or Identity	0.79
Threats to Survival or Physical Security	0.77
Vilification/Villainization	0.79
Explicit Dehumanization	0.74
None	0.73

Table 9: Inter-Annotator Agreement: Cohen's Kappa for Ukrainian war bloggers data.

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.80	0.90	0.91
Threats to Survival or Physical Security	0.76	0.81	0.82
Vilification/Villainization	0.78	0.89	0.88
Explicit Dehumanization	0.81	0.96	0.96
None	0.83	0.93	0.92

Table 10: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between majority vote and HQ-LLM (ChatGPT-4o) on Ukrainian war bloggers data.

Category	Instance Counts
Threats to Culture or Identity	120
Threats to Survival or Physical Security	74
Vilification/Villainization	136
Explicit Dehumanization	35
None	114
Total Data Points	**329**

Table 11: Summary of human annotations for Gab data.

Category	Cohen's
Threats to Culture or Identity	0.87
Threats to Survival or Physical Security	0.88
Vilification/Villainization	0.88
Explicit Dehumanization	0.92
None	0.91

Table 12: Inter-Annotator Agreement: Cohen's Kappa for Gab data.

	Dataset F1 Score		
Prompting Type	Russian	Ukrainian	Gab
No Additional Prompt	0.74	0.66	0.53
In-Context Learning	0.72	0.63	0.63
System Prompt	**0.78**	0.75	0.76
RDA	**0.78**	**0.76**	**0.77**

Table 13: F1-Score comparison across different prompting types and test sets (Russian Dataset, Ukrainian Dataset, Gab Dataset).

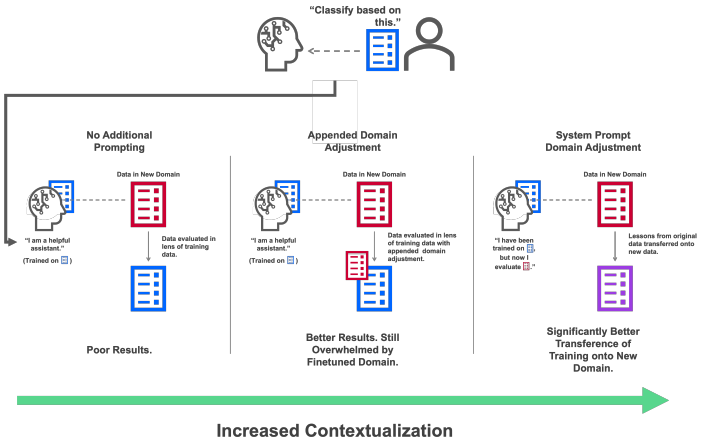


Figure 12: System Prompt Steering: demonstrates increased contextualization with the use of system prompt steering.

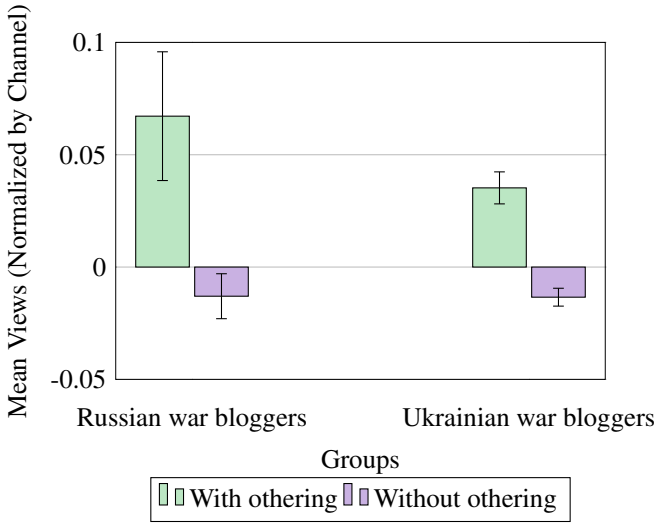
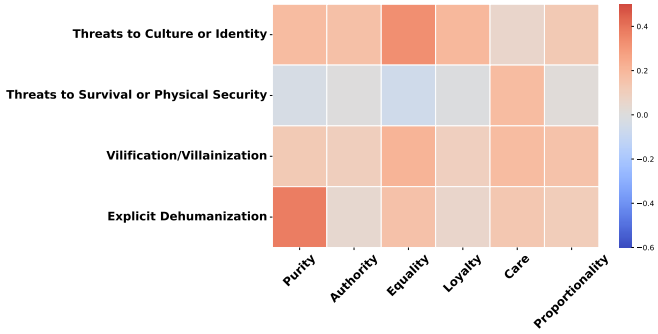
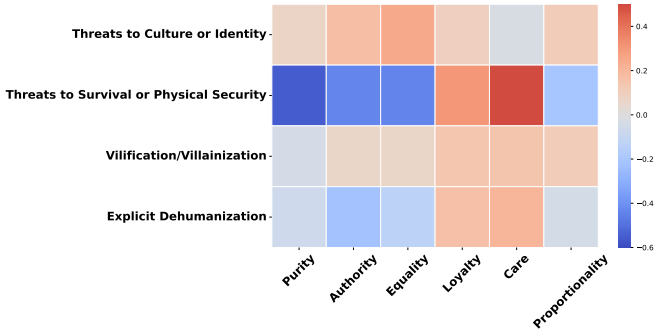


Figure 13: Comparison of mean views with and without othering (z-score channel-normalized) following crises. The bars represent the mean views for Russian and Ukrainian war bloggers, with and without othering, and the error bars indicate the standard error.



(a) Log-odds ratios for morality devices in **Russian war bloggers**' messages.



(b) Log-odds ratios for morality devices in **Ukrainian war bloggers**' messages.

Figure 14: Comparison of log-odds ratios for morality devices across othering categories in Russian and Ukrainian war bloggers' messages. The color intensity reflects the strength of the association, with warmer colors (red) indicating higher positive log-odds ratios and cooler colors (blue) representing negative or lower values.

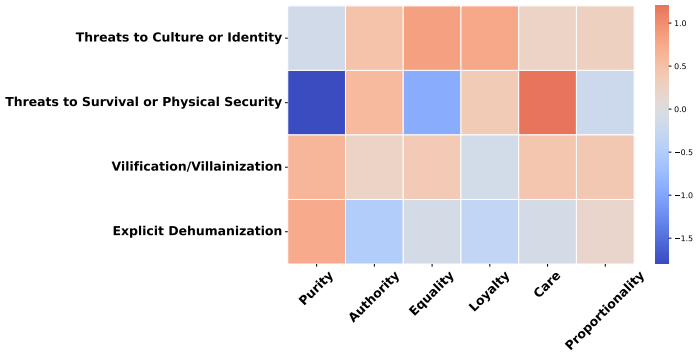


Figure 15: Heatmap displays the log-odds ratios for the use of various morality devices (Purity, Authority, Equality, Loyalty, Care, and Proportionality) across different othering categories (Threats to Culture or Identity, Threats to Survival or Physical Security, Vilification/Villainization, and Explicit Dehumanization) in the **Gab users'** messages. The color intensity reflects the strength of the association, with warmer colors (red) indicating higher positive log-odds ratios and cooler colors (blue) representing negative or lower values.

	Centrality Metric	
Community	Degree	Eigenvector
Russian	0.290 (+13.2%)	0.385 (+14.5%)
Ukrainian	0.177 (+32.1%)	0.136 (-7.8%)

Table 14: Centrality and othering messages following key events. Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities. All correlations are significant at the $p < 0.01$ level.

Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1a
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2a
2022-02-24	Russia invades Ukraine.	-
2022-04-19	Russia officially pivots to 'next phase' of war. Russia shifted its troops from the Kyiv offensive to Ukraine's eastern Donbas region, and the amassed forces launched a broad attack there on April 18. Ukraine called it a "new phase of the war."	3a
2022-06-01	The Biden administration authorizes an 11th presidential drawdown of security assistance to Ukraine valued at up to $700 million.	4a
2022-06-23	The Biden administration authorizes a 13th presidential drawdown of security assistance to Ukraine valued at up to $450 million.	5a
2022-07-08	The Biden administration announces $400 million in additional security assistance for Ukraine.	6a
2022-08-01	The Biden administration announces $550 million in additional security assistance for Ukraine.	7a
2022-08-01	United States Department of Defense announces approximately $1.1 billion in additional security assistance for Ukraine.	8a
2022-08-01	United States Department of Defense announces a significant new package of security assistance for Ukraine, including the authorization of a presidential drawdown of security assistance valued at up to $425 million, as well as $1.75 billion in Ukraine Security Assistance Initiative (USAI) funds.	9a

Table 15: Key events in the war discussed by Russian war bloggers. The "Key" column corresponds to the labeled vertical lines in Figure 5. Entries without a key were included in the data analysis but are not visualized due to their proximity to other points.

Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1b
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2b
2022-02-24	Russia invades Ukraine.	-
2022-03-02	Russia captures Kherson.	-
2022-03-21	Russian troops used stun grenades and gunfire to disperse a rally of pro-Ukrainian protesters in the occupied southern city of Kherson on Monday.	3b
2022-03-21	Russia abandons Kherson.	-
2022-04-01	Reports of Russian atrocities in Bucha begin to surface.	-
2022-07-03	Russia captures Lysychansk, all of Luhansk Oblast	4b
2022-08-29	Ukraine launches first major counteroffensive.	5b
2022-09-21	Ukraine forces Russian retreat.	6b
2022-11-11	Ukraine recaptures Kherson.	7b
2022-12-29	Major Russian missile attack on infrastructure facilities in Kyiv, Kharkiv, Lviv, and other cities.	8b
2023-02-09	Russia launches second spring offensive.	9b

Table 16: Key events in the war discussed by Ukrainian war Bloggers. The "Key" column corresponds to the labeled vertical lines in Figure 5. Entries without a key were included in the data analysis but are not visualized due to their proximity to other points.